

HABILITATION A DIRIGER DES RECHERCHES

Discipline : INFORMATIQUE

Title :

Toward generalizable Cross-Sensor AI Architectures: Theory, Applications and Future Directions

Yuta NAKANO

Acknowledgements

This document describes my research activities during the last 20 years.

(under editing... I complete at final step)

Abstract

Modern society is confronted with an unprecedented volume of data generated by diverse sensors embedded in smartphones, automobiles, and medical devices. Transforming these massive data streams into practical or meaningful information demands substantial expertise and manual effort. This habilitation thesis presents the author's research and development activities spanning more than two decades, unified by the overarching objective of automating meaningful information extraction from heterogeneous sensor data.

The research originates from the motivation to eliminate repetitive manual feature engineering and has evolved across three principal domains. In medical image analysis (Chapters 2–3), techniques for automatic algorithm construction based on genetic programming were proposed, covering organ segmentation from PET images, lesion extraction from CT images, and unsupervised Alzheimer's disease diagnosis using Self-Organizing Maps was also developed. Subsequent industrial research addressed clinically deployable technologies, including retinal layer segmentation from OCT images, CT-based texture analysis for virtual biopsy of Non-Alcoholic Steatohepatitis, and deep learning-based classification of pulmonary nodules. In the automotive domain (Chapter 4), deep neural network-based occupant detection using millimeter-wave radar was presented, and an event camera-based pedestrian detection framework was developed, achieving competitive performance with substantially reduced power consumption by using the sparsity of Spiking Neural Networks. Following these research and development activities, Chapter 5 outlines future research directions, with particular focus on multimodal data processing leveraging the asynchronous processing characteristics of Spiking Neural Networks — another advantage from their well-known energy efficiency.

This thesis presents my R&D trajectory from handcrafted feature design through deep learning to neuromorphic computing, and delineates my vision toward versatile, energy-efficient AI systems adaptable to diverse sensor modalities and real-world environments.

Keywords:

medical image analysis, genetic programming, image segmentation, deep learning, event camera, millimeter-wave radar, spiking neural network (SNN), neuromorphic computing, multi-modal data

Résumé (in French)

La société moderne est confrontée à un volume sans précédent de données générées par divers capteurs intégrés aux smartphones, aux véhicules et aux appareils médicaux. La transformation de ces flux de données massifs en informations exploitables ou pertinentes exige une expertise considérable et un travail manuel important. Cette thèse d'habilitation présente les activités de recherche et de développement menées par l'auteur depuis plus de deux décennies, unifiées par l'objectif global d'automatiser l'extraction d'informations pertinentes à partir de données de capteurs hétérogènes.

La recherche trouve son origine dans la volonté d'éliminer l'ingénierie manuelle répétitive des caractéristiques et s'est développée dans trois domaines principaux. En analyse d'images médicales (chapitres 2-3), des techniques de construction automatique d'algorithmes basées sur la programmation génétique ont été proposées, couvrant la segmentation d'organes à partir d'images TEP, l'extraction de lésions à partir d'images TDM, et un diagnostic non supervisé de la maladie d'Alzheimer utilisant des cartes auto-organisées a également été développé. Les recherches industrielles qui ont suivi ont porté sur des technologies cliniquement déployables, notamment la segmentation des couches rétiniennes à partir d'images OCT, l'analyse de texture basée sur la tomodesitométrie pour la biopsie virtuelle de la stéatohépatite non alcoolique, et la classification des nodules pulmonaires basée sur l'apprentissage profond. Dans le domaine automobile (chapitre 4), une détection des occupants basée sur des réseaux neuronaux profonds et utilisant un radar à ondes millimétriques a été présentée, et un cadre de détection des piétons basé sur une caméra événementielle a été développé, atteignant des performances compétitives avec une consommation d'énergie considérablement réduite grâce à la parcimonie des réseaux neuronaux à impulsions. À la suite de ces activités de recherche et développement, le chapitre 5 présente les orientations futures de la recherche, en mettant particulièrement l'accent sur le traitement des données multimodales tirant parti des caractéristiques de traitement asynchrone des réseaux neuronaux à impulsions — un autre avantage découlant de leur efficacité énergétique bien connue.

Cette thèse présente mon parcours en R&D, de la conception artisanale de caractéristiques au calcul neuromorphique en passant par l'apprentissage profond, et expose ma vision de systèmes d'IA polyvalents et économes en énergie, adaptables à diverses modalités de capteurs et à des environnements réels.

Mots-clés

analyse d'images médicales, programmation génétique, segmentation d'images, apprentissage profond, caméra événementielle, radar à ondes millimétriques, réseau neuronal à impulsions (SNN), calcul neuromorphique, données multimodales

Contents

Table of contents

Contents

Acknowledgements	2
Abstract	3
Résumé (in French)	4
Contents	5
List of Figures.....	9
List of tables.....	12
Chapter 1 Introduction	13
1.1. Motivation	13
1.2. Main contributions and overview	14
Chapter 2 Evolutionary Image Processing for medical images	17
2.1. Approaches for constructing medical image processing algorithm	17
2.1.1. Background	17
2.1.2. Objective	18
2.1.3. Prior arts in medical image processing	20
2.1.4. “Evolutionary image processing” approach.....	21
2.1.5. How to apply evolutionary image processing to medical image data	26
2.2. 3D-ACTIT	29
2.2.1. Overview	29
2.2.2. Input of three-dimensional training image set	30
2.2.3. Fitness function	30
2.2.4. Three dimensional of image processing filters	31
2.2.5. Three dimensional medical images used in experiments	32
2.2.6. Experiments on organ segmentation from PET images	35
2.2.7. Experiments by using 3D-ACTIT with filters using target regional features	43
2.2.8. Summary.....	59
2.3. Automatic image processing filter generation.....	61
2.3.1. Overview	61
2.3.2. Feature calculation in sliding window.....	62

2.3.3.	Image transformation using calculated features	63
2.3.4.	Integration of automatic image processing filter generation into ACTIT	64
2.3.5.	Integration of automatic image processing filter generation into FFGIN.....	73
2.3.6.	Summary.....	80
2.4.	Tissue classification using local features.....	81
2.4.1.	Overview	81
2.4.2.	Image diagnosis for Alzheimer’s disease	81
2.4.3.	Tissue classification using self-organizing maps.....	82
2.4.4.	Experiment of tissue classification on brain PET image	85
2.4.5.	Tissue classification experiment using CT image	88
2.4.6.	Summary.....	90
Chapter 3 Precise Contour Detection and Anormal Pattern Recognition in Medical domain		92
2.1.	Introduction	92
3.2.	Retinal layer segmentation in Optical Coherence Tomography (OCT)	93
3.2.1.	Background	93
3.2.2.	Proposal method	94
3.3.	Organ segmentation in CT	95
3.3.1	Background	95
3.3.2	CT phase estimation	96
3.3.3	Refinement using Graph cut	97
3.6.	Lymph node segmentation in CT	99
3.4.1	Background	99
3.4.2	Level set with scale and shape normalization.....	100
3.6.	Non-Alcoholic Steatohepatitis (NASH) assessment based on texture analysis of liver regions in CT images	102
3.5.1.	Background	102
3.5.2.	Issues and challenges.....	102
3.5.3.	GLCM based quantitative biomarker	103
3.5.4.	Experimental results	104
3.5.5.	Conclusion	105
3.6.	Benign/malignant classification of lung nodules using deep learning.....	106
3.6.1.	Background	106

3.6.2.	Issues and challenges.....	106
3.6.3.	Technical contributions.....	107
3.7.	Summary.....	107
Chapter 4	Sensor data processing with AI in Automotive domain	109
4.1.	Introduction	109
4.2.	Radar signal processing for passenger detection in vehicle	110
4.2.1.	Background	110
4.2.2.	Conventional methods.....	111
4.2.3.	Technical contributions.....	112
4.3.	Pedestrian detection on event camera data.....	116
4.3.1.	Background	116
4.3.2.	Conventional methods.....	117
4.3.3.	Proposed method	118
4.3.4.	Experimental results	120
4.4.	Feasibility study of ANN2SNN.....	122
4.4.1.	Our mission and objectives.....	122
4.4.2.	Technical background	123
4.4.3.	Scientific contributions	125
4.5.	Summary.....	132
Chapter 5	Future plan.....	134
5.1.	Introduction	134
5.2.	SNN object detection with ultra-low energy consumption	134
5.2.1.	Background and conventional methods	135
5.2.2.	Our current achievements	135
5.2.3.	Future plan	141
5.3.	Multi-modal SNN: handling various sensor data simultaneously.....	141
5.3.1.	Prior arts of SNN for various modalities	141
5.3.2.	Future plan	142
Chapter 6	Curriculum Vitae	144
	General details.....	144
	Career	144
	Research directions	144

Teaching activities / PhD students supervision.....	144
Publications	145
References	148
Appendix A Automatic construction of video image processing using 3D-ACTIT	154
A.1. Summary.....	154
A.2. Training dataset	154
A.3. Moving object extraction from video data	155

List of Figures

Figure 1 Processing flow of ACTIT.....	23
Figure 2 Structure of FFGIN	26
Figure 3 Construction of 3D voxel data from multi-slice images	30
Figure 4 Examples of PET image	34
Figure 5 Examples of diffusion-weighted image	35
Figure 6 Training data set for extracting kidney and lung regions	36
Figure 7 Results on training data for kidney region extraction	38
Figure 8 Obtained tree-structural filter in kidney extraction experiment.....	39
Figure 9 Results on training data for lung region extraction.....	40
Figure 10 Obtained tree-structural filter in lung extraction experiment	41
Figure 11 Results on testing data.....	43
Figure 12 Training data set for extracting liver regions.....	45
Figure 13 Results on training data for liver region extraction.....	46
Figure 14 Obtained tree-structural filter in liver extraction experiment	47
Figure 15 Transients of fitness with and without liver feature	49
Figure 16 Results on testing data for liver region extraction	51
Figure 17 Artifact and abnormal signals in diffusion-weighted image.....	52
Figure 18 Training dataset for extracting abnormal signals from diffusion-weighted image.	54
Figure 19 Results on training data for abnormal signal extraction on diffusion-weighted image	55
Figure 20 Results on testing data.....	57
Figure 21 Obtained tree-structural filter in liver extraction experiment	59
Figure 22 Feature calculation using sliding window	63
Figure 23 Image processing filter considering features	64
Figure 24 Training data for extracting lung region from PET images	65
Figure 25 Experimental result using conventional ACTIT.....	67
Figure 26 Experimental result of ACTIT using proposed filter generation method	69
Figure 27 Results on testing images	70
Figure 28 Training data for extracting liver region from CT image.....	71
Figure 29 Result on training data for liver segmentation by using automatic filter generation method	72
Figure 30 Genotype and phenotype in PT-FFGIN.....	73
Figure 31 Training data for extracting lung, heart and entire body	75
Figure 32 Training data for extracting lung, cancer region and entire body	75
Figure 33 Analysis of obtained network structured filter for segmentation of lung, heart and entire body	77
Figure 34 Results on testing data.....	78
Figure 35 Analysis of obtained network structured filter for segmentation of lung, cancer region and entire body	79

Figure 36 Results on testing data.....	80
Figure 37 Examples of brain PET images of healthy individuals and Alzheimer’s disease patients.....	82
Figure 38 Calculation of texture features using sliding window	83
Figure 39 Algorithm of self-organizing map.....	84
Figure 40 Tissue classification using self-organizing map	85
Figure 41 Features used in experiments.....	86
Figure 42 Brain PET tissue classification results.....	87
Figure 43 Application to testing images.....	88
Figure 44 Tissue classification of CT image.....	90
Figure 45 An example of OCT and boundaries of retinal layers.....	94
Figure 46 Profiles made by intensity information on OCT	95
Figure 47 Phase invariant points in CT for image registration	97
Figure 48 Blob shape conversion by using LoG.....	101
Figure 49 How to compute textural signature	104
Figure 50 METAVIR score classification by using textural signature	105
Figure 51 Conventional method and our idea explained in this section.....	111
Figure 52 Pipeline of method using deep learning with distance map.....	112
Figure 53 Raw milli-wave sensor data composed of complex numbers	113
Figure 54 Architecture of our neural network which input raw milli-wave sensor data.....	114
Figure 55 Data distribution in each class in training dataset	115
Figure 56 Data distribution in each class in testing dataset.....	115
Figure 57 Passenger detection results in each method	116
Figure 58 Examples of images reconstructed from event data.....	117
Figure 59 Examples of object detection on event data with and without using patches	119
Figure 60 Data representation as image by using binary, timestamp and count.....	120
Figure 61 Detection performance “morphology based VS patch based”	121
Figure 62 Detection performance between data representations	121
Figure 63 Number of FP (False Positive) in two data representations.....	122
Figure 64 Our mission in this project and issue to which we addressed	123
Figure 65 Difference between CNN and SNN.....	124
Figure 66 Processing flow for training and inference in hand gesture recognition system .	128
Figure 67 Prepared CNN architectures for performance shift investigation.....	129
Figure 68 (left) Temporal evolution of events. Positive events (new changes) are in white and negative (disappearing) ones are in black. (right) Count-image of accumulated events within 5ms.	130
Figure 69 Boxplot for performance shift after SNN conversion and chart describing gesture recognition accuracy of CNN and SNN	131
Figure 70 Standard deviation of gesture recognition accuracy in each architecture.....	131
Figure 71 SNN architecture for pedestrian detection	137
Figure 72 Preprocessing in our lane detection	140

Figure 73 Idea of multi-modal SNN	143
Figure 74 Training image set for moving object extraction.....	155
Figure 75 Results on training data for moving object extraction.....	156
Figure 76 Results on testing data.....	157

List of tables

Table 1	Image processing filters installed in ACTIT	24
Table 2	Image processing filters installed in 3D-ACTIT.....	31
Table 3	GP parameters in 3D-ACTIT	36
Table 4	Additional 3D image processing filters for liver segmentation.....	44
Table 5	Additional 3D image processing filters for asymmetry signal extraction	53
Table 6	Image processing filters considering features and their symbols in ACTIT.....	66
Table 7	Image processing filters integrated into PT-FFGIN.....	74
Table 8	Comparison between CNN and SNN	128
Table 9	Ablation study of our weight term in both the SNN and equivalent ANN version of our network.....	138

Chapter 1

Introduction

1.1. Motivation

Modern society is inundated with an unprecedented volume of imaging technologies (visual data) and sensor data. From smartphones and vehicles to medical environments, a wide variety of sensors are continuously operating, generating large-scale, high-resolution data streams. While such data have greatly enriched our daily lives and served as a foundation for safety, transforming these massive raw data into meaningful information for human decision-making requires significant time and effort.

For instance, in the automotive domain, data obtained from multiple onboard sensors must be processed rapidly and accurately to detect obstacles or hazards, and to trigger actions such as braking. In the medical domain, a single patient may generate thousands of medical images, all of which must be examined comprehensively before a final diagnostic decision can be made—requiring substantial expertise and effort.

To alleviate such processing burdens, AI technologies have advanced rapidly. However, prior to the rise of deep learning, Feature Engineering played a central role. In this paradigm, the problem at hand had to be carefully analyzed, effective features identified, and extraction algorithms implemented—tasks that were entirely manual. In medical image diagnosis, for example, feature engineering was performed separately for each image modality, disease type, and sometimes even for individual patients. Subsequently, the choice of classifier—whether a neural network, support vector machine, or k-means clustering—depended on the specific situation. As a result, researchers and medical professionals were burdened with extensive pre-processing work.

Motivated to put an end to this repetitive feature engineering process, the author focused on automatic construction of image processing algorithms as the central theme of research. From 2003 as a Master’s student, and from 2005 as a PhD student at Yokohama National University in Japan, research was conducted primarily in the medical field, with a focus on developing algorithms for automated medical image analysis to support radiologists in diagnostic tasks (see Chapter 2).

After joining Canon Inc. in 2008, then MEDIAN Technologies in 2016, the author worked on AI technologies for medical and pharmaceutical processes, aiming to achieve both high accuracy and robustness while improving workflow efficiency—key performance indicators for product-level deployment. After that, engaging in the application of deep learning to medical practice, collaborating with numerous hospitals to secure large-scale training datasets and bridging the gap toward practical implementation (see Chapter 3).

In 2018, the author moved to MRA EUROPE, shifting the research focus to the automotive industry. This work includes the processing of on-vehicle sensor data and the development of sensing AI for production sites. Addition to that, it provides opportunities to work with a variety of sensor modalities, including next-generation devices such as event cameras (see Chapter 4).

Based on these experiences, the author intends to continue advancing research and development of versatile AI systems capable of adapting to diverse sensor data and applications. In particular, future work will explore multimodal signal processing leveraging the asynchrony of Spiking Neural Networks (SNNs) and applying these technologies to high-precision robotic control in the automotive industry. This habilitation thesis will present an integrated discussion of the evolution of these foundational technologies, their real-world implementations, and the prospective research directions.

1.2. Main contributions and overview

The specific research activities and contributions are described in detail in the subsequent chapters.

Chapter 2 presents a research and development effort aimed at addressing challenges in medical imaging diagnosis, particularly through the use of a technique that employs genetic programming—one of the evolutionary computation methods—to assemble image-processing filters in a tree structure, much like building blocks.

In Section 2.2, this approach is applied to the automatic construction of complex three-dimensional image-processing algorithms. The conducted experiment focused on organ segmentation from PET (Positron Emission Tomography) images with ambiguous boundaries that is a task for which manual algorithm design is particularly difficult. The study evaluated how AI could generate viable solutions for such challenging problems.

Section 2.3 introduces an idea to improve the image-processing filters themselves. Traditionally, the approach relied on pre-implemented image transformation algorithms with parameters fixed in advance. In contrast, the author’s proposal incorporates a mechanism to adapt the transformation algorithm according to the characteristics of the input image. This not only improved accuracy but also enhanced robustness against testing images (which is not used as training data), thanks to processing that better captured image-specific features. The performances of this method are shown by applying to the PET and CT images in this section.

Furthermore, Section 2.4 presents a diagnostic support method based on unsupervised learning using a Self-Organizing Map (SOM). While the evolutionary computation-based approach required training data containing explicit ground-truth labels, many scenarios in medical diagnosis involve detecting subtle abnormalities that deviate from normal patterns, without a clearly defined correct example. To address this, the author developed a method

for unsupervised clustering of local image features and applied it to Alzheimer's disease diagnosis using brain PET images.

Then, Chapter 3 shows my achievements conducted in R&D department of private companies (CANON Inc. and MEDIAN Technologies). In these companies, I continued to work in medical domain, however, I focused on rather developing practical technologies than the technologies explained in Chapter 2.

Section 3.2 presents retinal payer segmentation technologies on OCT (Optical Coherence Tomography) data. As OCT is very noisy, I employed local image features to emphasize the borders of retinal payers, and determined the borders by using profile patterns.

Then, Section 3.3 and 3.4 introduce organs and lesions segmentation technologies. By using Graph cut and level set methods, we succeeded precise contour determination.

In Section 3.5, texture analysis on CT for virtual biopsy is explained. We gather a lot of data which includes liver disease named Non-Alcoholic Steatohepatitis (NASH) and evaluated if our methodology is useful.

After R&Ds written in section 3.5, the epoch of deep learning began. Therefore, I started to use it for benign/malignant classification of lung nodule, which is shown in Section 3.6.

From Chapter 4, I moved my field to automotive domain, and continue research activities. Through the work written in Section 3.6, I really understood the potential of deep neural network and also its difficulties such as preparation of huge training data. In this sense, automotive domain is easier to obtain large amount of data than medical field, therefore, we can have more chances to access cutting edge technologies, that is one of the reason why I change the domain.

In the sections in this chapter, I always use deep learning technology. For instance, section 4.2 introduce radar (milli-wave) signal processing for passenger detection in vehicle by using deep neural network handling complex number.

Then, our team started the project to develop "super low energy consumption system" due to the waves of edge computing, electric vehicle, etc. Therefore, we tested event sensor (dynamic vision sensor) who captures only intensity changes and whose energy consumption is lower than general imager such as RGB camera. As the data format of event sensor is completely different, we faced many issues. However, we achieved finally competitive performance in pedestrian detection, explained in Section 4.3.

After feasibility study of event sensor, we moved to another energy efficient technologies, to be more specific, Spiking Neural Network (SNN). Aside from sensor part, we tried to reduce energy consumption also in algorithm part. As a first trial, in section 4.4, we tested ANN2SNN which converts trained conventional neural network to SNN. In this study, I observed energy reduction and performance shift (decrease) by conversion.

Through above activities, we set a direction that we are going to target SNN technology to make our core competence. Therefore, we are now focusing to develop cutting edge algorithm integrated into SNN. Chapter 5 introduces our recent achievements in this domain and show briefly our future direction.

Finally, an extended curriculum vitae is provided in Chapter6.

Chapter 2

Evolutionary Image Processing for medical images

2.1. Approaches for constructing medical image processing algorithm

2.1.1. Background

In recent years, new technologies such as electronic medical records and teleconferencing systems have been introduced into clinical settings, and the role of information processing technology in medicine has become increasingly important. Among these, the development of image diagnosis technology has been remarkable, with imaging devices such as CT and MRI becoming increasingly high-resolution and high-speed. For example, with multi-slice CT, it is now possible to acquire more than a thousand images from a single patient in a matter of seconds. Additionally, technologies such as Positron Emission Tomography (PET) and fMRI, which visualise bodily functions, have been developed, enabling more detailed diagnostic imaging by leveraging the unique characteristics of each image. Currently, PET images acquired simultaneously with CT are used to diagnose from both morphological and functional perspectives, and comprehensive diagnoses are made by combining multiple types of images. However, the increasing volume of information resulting from handling large numbers of images of various types, as well as the need to adapt to newly developed diagnostic technologies, has undeniably increased the burden on physicians. There is a growing demand for diagnostic support systems (Computer-Aided Diagnosis; CAD) that can effectively utilise large amounts of information with reducing the burden.

When considering diagnostic support systems, image processing technology is indispensable, and as medical imaging devices continue to advance, the demand for image processing and recognition technology is growing stronger. To date, many medical image processing algorithms have been proposed, and their effectiveness has been demonstrated in various images and cases. However, many of these image processing algorithms are limited to specific types of images or cases, and when the target changes, adjustments to the processing procedures or parameters are necessary. This adjustment requires knowledge and experience in image processing, as well as specialised knowledge in the field of the target images, meaning that experts must trial and error each time depending on the images or cases being handled.

To address this issue, research on diagnostic support systems using image processing expert systems and probabilistic atlases of the human body has been actively conducted. These studies create databases of previously established processing procedures, parameters, and images, and use them as references to construct image processing algorithms and create human body maps. By employing these methods, image processing can be constructed

without requiring specialised knowledge when dealing with problems similar to those encountered in the past. However, these methods are difficult to apply to images or special cases that have not been handled in the past, as they require a historical database and probabilistic atlas that have not been constructed.

The laboratory that I was belonging to is conducting research on the automatic construction of image processing and recognition algorithms. One of these is a tree-structured image transformation automatic generation system called ACTIT (Automatic Construction of Tree-Structural Image Transformation), which combines image processing filters in a tree-like structure. This system has been applied to various problems and has demonstrated its effectiveness. ACTIT requires only image-based training data and does not necessitate prior knowledge of image processing or databases. It constructs tree-structured image processing filters by combining pre-defined filters using genetic programming (GP). This approach eliminates the developer's burden, which was heavily dependent on knowledge and experience in traditional image processing methods, as well as the limitations of knowledge-based processing procedures based on expert systems and the user's workload caused by system interaction. Additionally, other evolutionary image processing methods have been proposed, enabling image recognition and classification without limiting the types of images handled. These methods have been applied to various problems and demonstrated their effectiveness. However, when considering their application to medical image processing, it is believed that installing the unique characteristics of medical images into this philosophy can lead to the development of more advanced image processing and recognition algorithms.

2.1.2. Objective

In this section, I focus on evolutionary image processing for medical image processing. Medical image processing requires high-dimensional complex algorithms and image analysis and conversion that consider the pixel value characteristics of each tissue. Therefore, we propose (1) automatic construction of high-dimensional image processing by combining image processing filters, (2) automatic generation of image processing filters considering local features, and (3) evolutionary medical image processing using local features for tissue classification.

(1) Automatic construction of high-dimensional image processing

Recent advances in imaging devices have increased the use of 3D voxel and video data in medical imaging. While this high-dimensional image data enables more detailed diagnosis, it also increases the number of images to be handled, placing a heavy burden on doctors. To achieve detailed diagnoses while reducing the burden on physicians, research on high-dimensional image processing is actively underway. However, as the dimension increases, the number of information and parameters to consider in the image space grows, leading to more complex algorithms. Therefore, we propose 3D-ACTIT, which applies the image processing automatic construction method of ACTIT to 3D image processing. In 3D-ACTIT, we have modified ACTIT to enable the merging of multiple cross-sectional images and

their handling as voxel data, and by fully 3D-ising all filters, we have made it possible to automatically construct 3D image processing. We will attempt to automatically construct segmentation processing for each organ in 3D-PET and processing to extract abnormal signal regions from diffusion weighted images using this 3D-ACTIT. For example, PET images do not exhibit consistent grayscale values even within the same organ or tissue, and their boundaries are often ambiguous, making them challenging for conventional image processing algorithms. We will verify whether 3D-ACTIT can automatically construct complex image processing by simply providing training data.

(2) Automatic generation of image processing filters considering image features

When making a diagnosis, doctors judge not only the signal intensity in an image but also the signal patterns and texture to determine what kind of tissue a certain area is and whether it is normal or abnormal. When handling medical images, it is extremely important to determine which tissue a certain point or area belongs to. In medical image processing as well, it is desirable to convert images while being conscious of which tissue each area belongs to. Therefore, we propose a method that considers the characteristics of each organ and tissue in medical images by calculating local texture features from the images, capturing the characteristics of each region, and performing image conversion according to those characteristics. In this method, we first generate feature images that represent the extent to which each feature is present in which part of the image. Based on these feature images, we automatically generate image processing filters that apply transformations such as strengthening or weakening the parts where features are strongly present. We introduce this method into filter combination methods such as ACTIT and aim to construct a highly advanced and versatile algorithm by combining the proposed filters.

(3) Tissue classification using local features

Medical image diagnosis requires not only detecting abnormal areas such as malignant tumours but also analysing each tissue to determine functional decline by comparing with normal patterns in healthy individuals and identifying abnormal areas relative to other tissues based on each patient's overall characteristics. Physicians currently perform tissue analysis based on their mental databases, which are prone to subjectivity and heavily dependent on experience. There is a need for a method that objectively analyses each tissue by capturing image characteristics. To realise this objective image analysis method, we propose a tissue classification method using local features. This method clusters pixels based on calculated local texture features and classifies them into homogeneous tissues or normal/abnormal tissues. In the latter section, we adopt self-organising maps as the clustering method and map the calculated multidimensional local texture features onto a two-dimensional map to determine whether they belong to the same tissue type. To demonstrate the effectiveness of this method, we applied it to brain PET images of Alzheimer's disease patients, classified each tissue type, and attempted to identify regions with impaired brain function.

In the latter sections present above proposed methods, experimental results, and confirms their effectiveness.

Structure of this section

This chapter is structured as follows. First, section 1.2 introduces previous studies on medical image processing. Next, section 1.3 explains the 3D-ACTIT system, which is a 3D version of ACTIT, and describes studies applying 3DACTIT to medical images. section 1.4 explains a method for automatically generating image processing filters considering features, and presents experimental results of applying the implemented filter combination method to medical images. section 1.5 describes a tissue classification method using local features, and presents results of analysing brain PET images obtained from Alzheimer's patients and abdominal CT images using this method.

2.1.3. Prior arts in medical image processing

Image processing technology is used in various fields and is widely present in our daily lives. To date, numerous image processing algorithms have been proposed, and their effectiveness has been demonstrated [1][2]. However, with the recent advancements in imaging devices and network technologies enabling high-speed data communication, the opportunities to handle image data have increased, leading to higher demands on image processing technology. In particular, in the medical field, the high resolution and high speed of imaging devices such as CT and MRI used for diagnosis have enabled the acquisition of large amounts of data from patients, making it possible to perform more detailed image diagnosis. However, the increasing amount of data is placing a heavy burden on doctors who use it for diagnosis. In addition, the types of images used for diagnosis, such as PET, diffusion-weighted images, and fMRI, are increasing, and the demand for medical image processing technology is growing year by year. To address these challenges, research on diagnostic support systems and expert systems that utilise image processing to reduce number of tasks on physicians, extract more information from images, and enable more detailed diagnoses is actively underway [3][4]. By employing image processing technology, it is possible to evaluate the progression of symptoms and other factors in a quantitative and objective manner, free from subjective interpretation. This section introduces previous research on medical image processing.

Traditional image processing approaches

In diagnostic support systems, image transformation techniques are used to highlight abnormal areas for physicians or to detect abnormalities that are too small to be detected by the human eye. Research on image transformation for organ segmentation and enhancement of abnormal signals has been actively conducted, and its effectiveness has been demonstrated [5][6][7]. Kobata et al. proposed a vector concentration filter that uses brightness gradients to emphasise shadows with very weak contrast, and demonstrated its effectiveness through application to various medical images. This vector concentration filter prepares gradient patterns such as points, lines, and surfaces, enabling the emphasis of shadows with shapes matching each pattern. Additionally, Shimizu et al. considered directional differences to

emphasise block-like shadows in chest X-ray images and demonstrated its effectiveness. In medical images, abnormal signals often have very weak contrast or share the same characteristics as normal signals, leading to their being obscured by other signals in simple thresholding processes. However, image transformation can be used to extract buried information.

Approaches using knowledge database

Physicians diagnose by comparing images with mental maps of the human body and normal patterns. By realising probabilistic atlases of the human body and databases of normal patterns on a computer, it is possible to use them as clues for organ extraction from human images or to identify abnormal areas by matching them with normal patterns[8]. In the study by Shimizu et al., an atlas was constructed, and multiple organs were segmented from abdominal 3D CT images containing various mixed tissues. In addition to databases of image information, research is also being conducted to construct knowledge and history databases of image processing procedures. Many image processing algorithms depend on the processing target, and when the target changes, the image processing procedures and parameters must be adjusted accordingly. In order to obtain the desired image processing, it is necessary to construct algorithms through trial and error based on the knowledge and experience of experts. To address this issue, methods for constructing general-purpose image processing algorithms capable of handling various images and requirements have been proposed [9][10][11]. These studies determine processing procedures by utilising a knowledge database of past history and image processing based on the characteristics of the target images. The knowledge database contains accumulated processing histories and algorithms whose effectiveness has been confirmed, allowing the construction of reliable image processing algorithms by referencing them. These methods are effective when knowledge about previously solved problems, similar image processing tasks, or the target images is established, and they have resolved various issues to date.

2.1.4. “Evolutionary image processing” approach

Our research group has proposed evolutionary image processing, which formulates image processing as an optimisation problem and automatically constructs image processing, recognition, and analysis algorithms. In this section, we describe ACTIT and GIN, representative examples of evolutionary image processing, and discuss whether evolutionary image processing is effective in the field of medical image processing.

ACTIT

The Automatic Construction of Tree-structural Image Transformation (ACTIT) is a system that applies genetic programming (Genetic Programming; GP)[13][14], which was developed based on genetic algorithms (Genetic Algorithm; GA)[12], an evolutionary computation method, to image processing. [15][16]. The user prepares a target image by manually creating the original image to be processed and the ideal output image using a painting tool, and if necessary, a training image set consisting of a weight image that represents the importance

of each pixel in the image using the magnitude of the grayscale values. When this training image set is input into ACTIT, the pre-prepared 1-input 1-output and 2-input 1-output image processing filters are combined into a tree-like structure using GP. As shown in Figure 2.1, the original image is input from all leaf nodes, and each undergoes image transformation. These are then sequentially combined using the 2-input filter to obtain a single output image. By comparing this output image with the target image, the tree is evaluated, and by repeating the GP-based generation process, an effective tree-structured filter is obtained.

To date, the ACTIT system has automatically constructed image processing tasks such as removing handwritten characters from images and extracting facial regions from colour images, demonstrating its effectiveness. Furthermore, the ACTIT system has been applied to medical image processing, achieving tasks such as cell wall extraction from microscope images and lung region extraction from X-ray images. Figure 1 shows the processing flow of ACTIT. Initially, randomly generated tree-structured filters are subjected to repeated generations of selection, crossover, and elimination via GP to bring their outputs closer to the target image. The process ends when the fitness reaches the maximum value of 1.0 or the number of generations reaches the setting (predefined) value, and the tree-structured image transformation with the maximum fitness is output. If the generated tree-structured image transformation utilises the common features of the target image group to perform the desired processing accurately, the same processing should be possible for images not included in the training image set, i.e., testing images.

Fitness function

The evaluation function for fitness evaluation in ACTIT is shown below.

$$Fitness = \frac{1}{N} \sum_{n=1}^N \left\{ 1 - \frac{\sum_{i=1}^W \sum_{j=1}^H w_{ij}^n |o_{ij}^n - t_{ij}^n|}{V_{max} \sum_{i=1}^W \sum_{j=1}^H w_{ij}^n} \right\} \quad (1)$$

Here, n represents the n th image set, the target image is t_{ij}^n , the output image generated by the tree-structured image processing filter is o_{ij}^n , and the weight image representing the importance in the image is w_{ij}^n . Furthermore, in the i and j directions, the number of pixels be W and H . Here, to normalise the fitness value between 0.0 and 1.0, we divide it by V_{max} , which is the maximum grayscale value in the image file format being processed. As can be seen from the equation, as the output image becomes more similar to the target image, the fitness value increases, and when they are completely identical, the fitness value becomes 1.0. When using a weight image, each pixel is weighted and the weight is reflected in the fitness value.

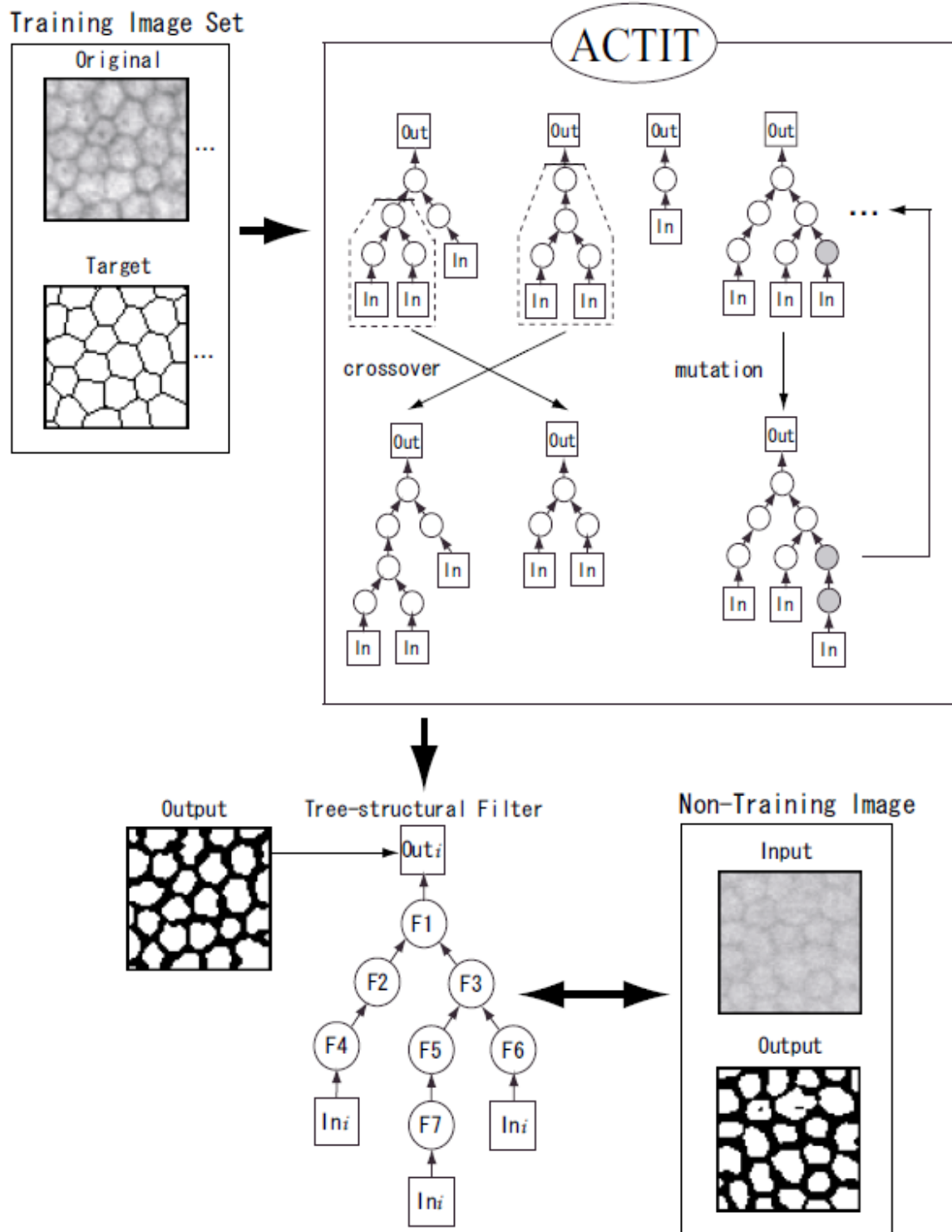


Figure 1 Processing flow of ACTIT

Image processing filters prepared in ACTIT system

The image processing filters provided in ACTIT are shown in Table 1. Conventional ACTIT achieves complex processing by combining simple processes such as filters that apply uniform pixel value conversion to images using a threshold value, neighbourhood processing filters,

and differential filters. Threshold values for processes such as binarization are determined based on the average pixel value of the entire image or by using a discrimination analysis method. The white-black edge enhancement filter is a filter that emphasises black edge areas extracted by the Laplacian filter in white and white edge areas in black.

Table 1 Image processing filters installed in ACTIT

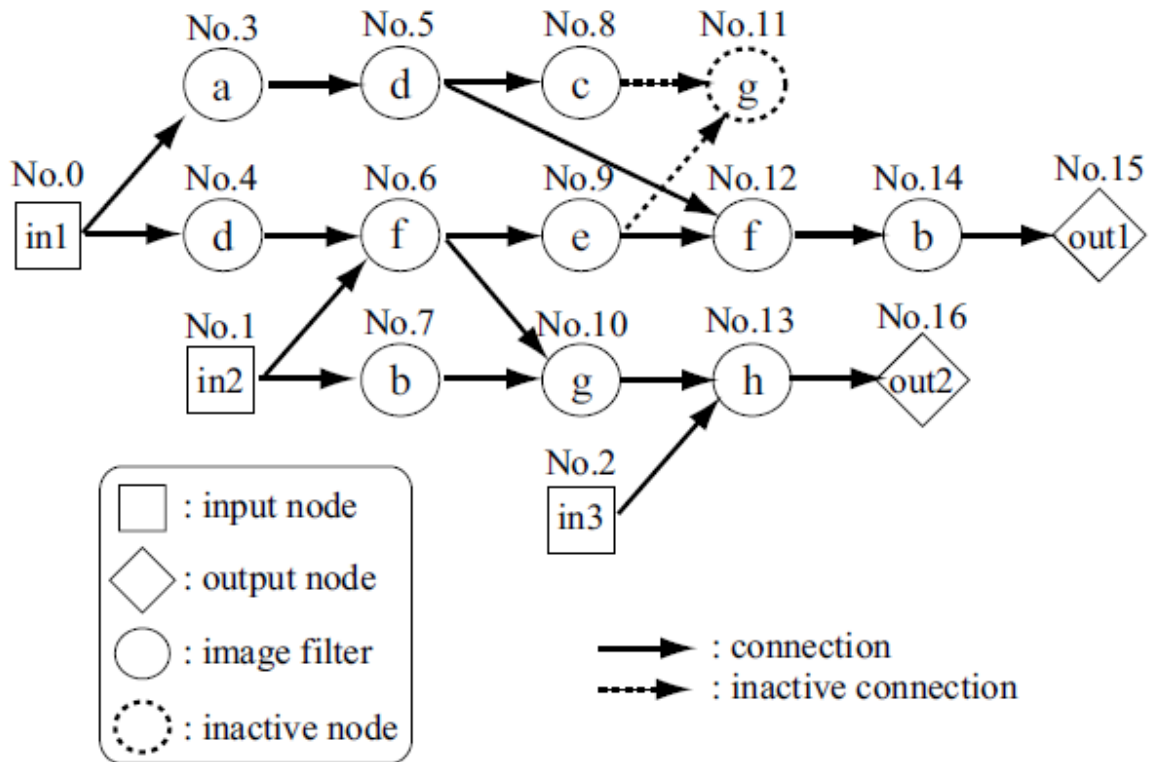
1-Input / 1-Output Filters	Description
Mea	Mean filter
Max	Maximum filter
Min	Minimum filter
Sob	Sobel filter
Lap	Laplacian filter
LiE	White edge enhancement
DaE	Black edge enhancement
LiP	Mean threshold: below mean $\rightarrow$ black
DaP	Mean threshold: above mean $\rightarrow$ white
LAB, LAW	Region with area smaller than threshold $\rightarrow$ white, else black
SAB, SAW	Region with area larger than threshold $\rightarrow$ white, else black
Inv	Inversion filter
LTr	Linear transformation filter
Gam	Gamma correction filter ($\gamma = 2$)
Squ	Squaring transformation filter
Exp	Expansion filter
Thi	Shrinking filter
BiM	Binary filter (mean threshold)
BDA	Binary filter (discriminant analysis)
2-Input / 1-Output Filters	Description
LoS	Logical AND: $\min(f_1, f_2)$
LoP	Logical OR: $\max(f_1, f_2)$
AlS	Arithmetic sum: $f_1 + f_2$
AlP	Arithmetic product: $f_1 \times f_2 / V_{max}$
BoS	Bounded sum: $f_1 + f_2 - V_{max}$
BoP	Bounded product: $\max(f_1 + f_2 - V_{max}, 0)$
DrS	Subtractive inhibition: $f_1 - f_2$ for $f_2 = 0$, else 0 (or $f_1, f_2 = 1$)
DrP	Inhibitory product: $f_1 - V_{max} \times f_2$ for $f_2 = V_{max}$, else 0
Dif	Difference filter: $ f_1 - f_2 $

GIN

ACTIT constructs complex image processing algorithms by assembling image processing filters into a tree structure. However, in addition to ACTIT, Genetic Image Network (GIN) has been proposed, which assembles image processing filters into a network structure with higher expressive power than a tree structure [17]. GIN handles network structures, therefore, it structurally includes tree structures and enables the expression of feedback and multiple outputs. In GIN, similar to ACTIT, a training image set consisting of original images and target images is prepared, and the network structure composed of image processing filters is evolved to bring the output closer to the target image. To achieve a structure with multiple outputs, the training image set includes the target image output multiple times, and the network structure is designed to have the same number of output nodes. The input original images are

processed for a predetermined number of steps and then output. Additionally, by modifying the network structure of GIN to a feedforward type, the Feed Forward Genetic Image Network (FFGIN) was proposed, enabling output without specifying the number of steps and allowing adjustment of parameters for each node. Its effectiveness has been demonstrated [18]. Figure 2.2 shows an example of the FFGIN structure. In the genotype, three elements are assigned to each node. The first represents the filter type (filter ID), and the remaining two represent the input nodes. If parameters are to be varied for each node, additional elements corresponding to the number of parameters are added. By representing the genotype in this manner, evolution via GA becomes possible. Furthermore, the approach of inputting a set of training images and adjusting filter combinations to approximate the target image is identical to ACTIT, enabling the sharing of fitness functions and predefined image processing filters. Using GIN, it is possible to represent the same problem using a compact network-structured algorithm rather than ACTIT's tree-structured algorithm. Additionally, as shown in Figure 2, it enables a single network-structured algorithm to simultaneously produce multiple outputs.

Phenotype



Genotype

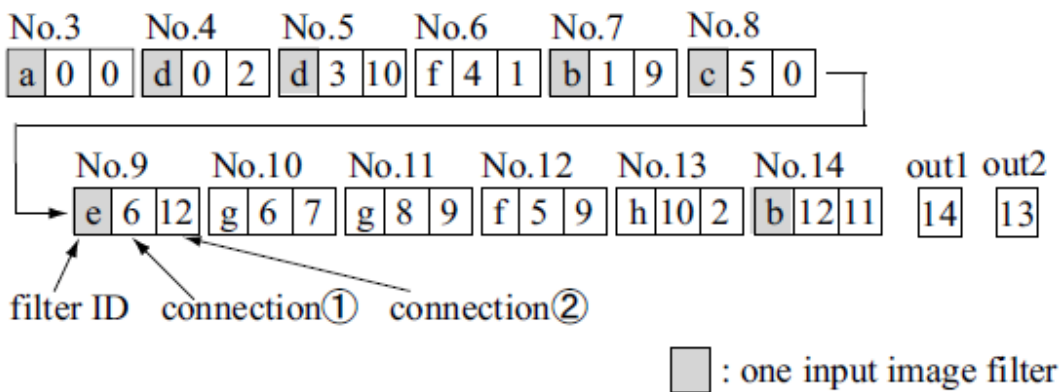


Figure 2 Structure of FFGIN

2.1.5. How to apply evolutionary image processing to medical image data

Conventional image processing algorithms are often specialised for specific images or problems, meaning that when the target changes, the processing steps or parameters must be adjusted. On the other hand, evolutionary image processing methods such as ACTIT and

GIN automatically construct algorithms by simply providing training data. Therefore, in medical image processing, which requires the extraction of various organs and tissues, and handles a wide variety of images and cases, evolutionary image processing methods are effective compared to traditional image processing methods, which require trial and error for each problem.

However, based on previous experiments, while evolutionary image processing can construct image processing algorithms for all types of images by simply providing training images, it is considered that by improving conventional evolutionary image processing, more advanced and robust image processing algorithms can be constructed in fields requiring specialised knowledge, such as medical image processing. Therefore, based on previous experiments and discussions with physicians, we propose an evolutionary “medical” image processing method that considers the following points:

- A. 3D image processing
- B. Feature-aware image transformation
- C. Feature-based image analysis

Recent advancements in imaging devices have made it possible to acquire 3D image data composed of multiple cross-sectional images and high-resolution, high-frame-rate video data, leading to increased opportunities to handle high-dimensional image data. Handling high-dimensional image data has enabled the discovery of cases that were previously undetectable and more detailed diagnoses. On the other hand, the amount of data used for diagnosis is directly proportional to the burden on physicians. Therefore, there is an increasing demand for both high-precision diagnosis and reduction of the burden on physicians. However, as the dimensionality of data increases, the number of information and parameters to be considered in the image space also increases, leading to more complex algorithms. Therefore, we propose 3D-ACTIT, a 3D version of ACTIT, by applying the concept of evolutionary image processing to achieve the automatic construction of 3D image processing. The description and experimental results of 3D-ACTIT are presented in next section.

Furthermore, several issues were identified from experiments with ACTIT and GIN. In automatic image processing experiments, while the output images from the training image set largely matched the expected results, applying the constructed algorithm to testing images often yielded results that did not match expectations due to minor differences in features. This was particularly evident in medical image processing, where minor differences in pixel values caused by individual variations frequently led to extraction errors. One of the reasons for these results is the image processing filter set provided by the system. The image processing filters provided by ACTIT and GIN are mostly simple algorithms that only consider threshold values calculated from the entire image and neighbouring pixel values, leading to the construction of algorithms that simply match the final output image to the target image. Additionally, when incorporating prior knowledge specific to the medical field that cannot be expressed using simple threshold determination methods, it is necessary to prepare image

processing filters on a case-by-case basis. To address this issue, we propose an image processing filter automatic generation method that automatically captures features within images and performs image transformations accordingly. This method automatically generates image processing filters from local features in images, eliminating the need to add filters specific to the target. Furthermore, by considering not only the intensity of pixel values but also texture features, it can capture the characteristics of each tissue in medical images, and is effective for testing images. An explanation of the image processing filter automatic generation method is provided in section 1.4.

In addition, there are problems that are difficult to address with supervised learning methods such as ACTIT and GIN. ACTIT and GIN are good at extracting regions with certain features by combining image processing filters, but when regions with features different from normal patterns appear in images due to, for example, functional decline in each organ or a specific case, it is difficult to extract them because explicit training data cannot be provided. A method that can objectively classify organs, tissues, normal tissues, and abnormal tissues from image features without providing training data is desired. Therefore, we propose a method for tissue classification by calculating local features of images and inputting them into a self-organising map using unsupervised learning. A self-organising map updates the map from the input data and maps multidimensional data to a two-dimensional map where data with similar features are grouped together. This method enables tissue classification even for images without explicit training data by analysing local features of the images. This tissue classification method is described in section 1.5. We propose evolutionary image processing that is applicable to medical images and aims to construct more advanced medical image processing algorithms, and demonstrate its effectiveness.

2.2. 3D-ACTIT

2.2.1. Overview

Recent advancements in medical imaging devices have made it possible to acquire high-dimensional image data such as 3D voxel data and dynamic images. For example, in image diagnosis using multi-slice CT, physicians reconstruct the acquired multiple cross-sectional images into 3D data in their minds and identify organs and determine normal or abnormal conditions by examining their spatial relationships. Currently, 3D viewers and diagnostic support systems that enable the reconstruction of 3D image data composed of multiple cross-sectional images and diagnosis from various angles have been developed. However, the high-resolution and high-dimensional nature of the image data being handled continues to advance, and new imaging technologies are being developed one after another, leading to increasing demands on image processing technology. To address these demands, research on 3D image processing is actively underway, and its effectiveness has been demonstrated [19][20]. However, in the field of medical image processing, where a wide variety of image types and cases exist, methods that require adjusting processing procedures and parameters each time the target changes are difficult to apply universally. Furthermore, 3D image processing algorithms cannot simply be adapted from existing 2D image processing algorithms, as increasing the number of dimensions requires consideration of additional information and parameters, generally leading to greater complexity. To address this issue, research is being conducted on expert systems that can automatically adjust processing procedures and parameters by leveraging knowledge and history databases related to 3D image processing [21][22]. These studies enable the automatic determination of processing procedures and parameters from a knowledge database by simply specifying the shape of the target object, and have been applied to CT images, demonstrating their effectiveness. However, when a knowledge database for the target images or cases is not established, it is difficult to apply this method to them.

Evolutionary image processing is a concrete implementation of the concept of "formulating image processing as an optimisation problem and solving it using evolutionary computation methods," and it is not limited to specific applications [23]. In this section, we propose 3D-ACTIT, an evolutionary image processing method that can automatically construct image processing without limiting the target, and apply it to medical image processing to achieve the automatic construction of complex 3D image processing. The 3D extension of ACTIT is achieved by storing the original image, which consists of multiple cross-sectional images, as 3D voxel data, preparing all filters as 3D image processing filters, and evolving a population of individuals consisting of randomly combined filters in a tree structure using a genetic programme (GP), thereby enabling the automatic acquisition of the required 3D image processing. Through this section, we propose 3D-ACTIT and demonstrate its effectiveness through experiments on 3D medical images such as PET and diffusion-weighted images.

2.2.2. Input of three-dimensional training image set

3D-ACTIT reads multiple cross-sectional images as a dataset and stores them as voxel data, as shown in Figure 3. By handling the data as voxel data, the depth direction can be considered, enabling edge detection that was not possible with 2D image processing during differential processing. Additionally, when handling human data consisting of multiple cross-sectional images such as CT or MRI, the threshold values used in image processing filters were previously calculated separately for each cross-sectional image in the 2D-ACTIT, resulting in significant dependence on the selected training image set. However, by treating the target region as a single 3D voxel data set, the threshold value can be calculated based on the entire target region.

For training data, we adopted a method where a single cross-sectional image and its corresponding target image, along with weight images (reference images where the importance in the target image is represented by grayscale values) if necessary, are prepared, and these are used as a training image set corresponding to the number of cross-sectional images. Here, multiple training image sets can also be used.

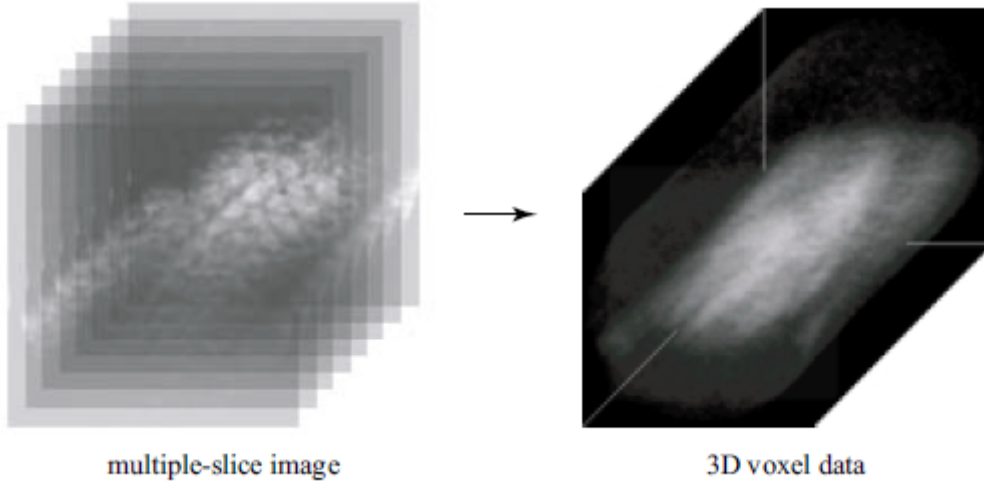


Figure 3 Construction of 3D voxel data from multi-slice images

2.2.3. Fitness function

The evaluation function for fitness evaluation, in 3D-ACTIT is shown below.

$$Fitness = \frac{1}{N} \sum_{n=1}^N \left\{ 1 - \frac{\sum_{i=1}^W \sum_{j=1}^H \sum_{k=1}^D w_{ij}^n |o_{ij}^n - t_{ij}^n|}{V_{max} \sum_{i=1}^W \sum_{j=1}^H \sum_{k=1}^D w_{ij}^n} \right\} \quad (2)$$

As in the conventional ACTIT, n represents the n th image set, and the target image is t_{ijk}^n , tree-structured image processing. The output image generated by the filter is denoted as o_{ijk}^n , and the weighted image representing the importance within the image is denoted as w_{ijk}^n . To handle 3D space, the number of pixels in the i, j , and k directions are denoted as W, H , and D , respectively, and the fitness is calculated. In 3D-ACTIT, the fitness also takes values from 0.0

to 1.0. As the image becomes more similar to the target image, the fitness increases, and when it perfectly matches the target image, the fitness becomes 1.0.

2.2.4. Three dimensional of image processing filters

In this section, experiments were conducted using basic 3-dimensional image processing filters as the basic filter set, as shown in Table 2. In the image processing filters in the table, the neighbourhood range is 26 neighbours, and the threshold is calculated using the average grey value of the entire image or discriminant analysis. In 3-dimensional image processing, in addition to 26 neighbours, 6 and 18 neighbours are also representative examples. For thresholding filters, methods that vary the threshold value depending on the image position are also considered, however for now, we use a simple algorithm with a neighbourhood range of 26 neighbours and a threshold value calculated using the average grey value and discriminant analysis to evaluate the basic performance of the system.

Table 2 Image processing filters installed in 3D-ACTIT

1-Input / 1-Output Filters	Description
-	Mean filter
M	Maximum filter
m	Minimum filter
d	Sobel filter
L	Laplacian filter
E	White edge enhancement
e	Black edge enhancement
j	Mean threshold: below mean $\rightarrow$ black
t	Mean threshold: above mean $\rightarrow$ white
i	Inversion filter
H	Linear transformation filter
h	Sharpening filter
J	Gamma correction filter ($\gamma = 2$)
j	Squaring transformation filter
G	Exponential transformation filter
g	Logarithmic transformation filter
N	Binary filter (mean threshold)
n	Binary filter (discriminant analysis)
2-Input / 1-Output Filters	Description
L	Logical AND: $\min(f_1, f_2)$
l	Logical OR: $\max(f_1, f_2)$
A	Arithmetic sum: $f_1 + f_2$
a	Arithmetic product: $f_1 \times f_2 / V_{max}$
B	Bounded sum: $f_1 + f_2 - V_{max}$
b	Bounded product: $\max(f_1 + f_2 - V_{max}, 0)$
U	Subtractive inhibition: $f_1 - f_2$ for $f_2 = 0$, else 0 (or $f_1, f_2 = 1$)
u	Inhibitory product: $f_1 - V_{max} \times f_2$ for $f_2 = V_{max}$, else 0
D	Difference filter: $ f_1 - f_2 $

2.2.5. Three dimensional medical images used in experiments

We applied the proposed 3D-ACTIT to 3-dimensional medical image data and attempted to automatically construct 3-dimensional image processing. In this section's experiments, we used PET images, which are gaining attention for their effectiveness in cancer detection and Alzheimer's disease diagnosis, and diffusion-weighted images, which are effective for diagnostic imaging of conditions such as cerebral infarction and cellular toxicity oedema. Both types of images do not directly represent the actual substances themselves but rather visualise metabolic processes or molecular movements in the human body, resulting in challenges such as ambiguous boundaries and the presence of artefacts in image processing.

Currently, new devices such as PET/CT, which can simultaneously acquire PET and CT images, have been developed and their effectiveness has been demonstrated [24][25]. PET/CT allows both PET and CT cross-sectional images to be acquired at the same location, enabling diagnosis using PET images with confirming the position using CT images. However, due to differences in spatial resolution and acquisition time, it is difficult to perfectly align PET and CT images. Additionally, in diffusion-weighted images, since they are acquired using MRI imaging equipment, MRI and diffusion-weighted images can be obtained at the same location. However, similar to PET, errors exist due to differences in spatial resolution and imaging time, and physicians must mentally align the images to make diagnoses. Therefore, studies evaluating absorption correction and scattering correction using respiratory synchronisation during imaging have been conducted, and their effectiveness has been demonstrated [26][27]. However, few studies have focused on processing PET or diffusion-weighted images themselves post-imaging, with most examples utilising functional information from PET or diffusion-weighted images based on morphological information from CT or MRI.

In this study, due to the increasing number of images read by physicians and concerns about radiation exposure, we developed an image processing method that uses only PET images or diffusion-weighted images, without relying on CT or MRI information. By focusing on PET images or diffusion-weighted images, we aim to reduce the burden on physicians and patients. Additionally, this approach allows for more accurate registration with CT or MRI images and enables the extraction of more information from functional images when assessing organ function, potentially leading to more detailed diagnoses. For the above reasons, we created a set of training images from PET images and diffusion-weighted images and input them into 3D-ACTIT. Specifically, we automatically constructed processes to extract organ regions from PET images and to extract lesion areas from diffusion-weighted images. First, PET images are explained in next subsection, then, diffusion-weighted images in the subsection after that. After the explanation of these images, experimental results using PET images and diffusion-weighted images are presented.

PET image

PET stands for Positron Emission Tomography. It involves administering a radioactive drug labelled with a radioactive isotope that emits positrons to a patient and then capturing the

radiation emitted from the body using a camera. By capturing the distribution of radiation as cross-sectional images, this imaging technique produces visual representations of organ function. The most commonly used radiopharmaceutical in PET is fluorodeoxyglucose (FDG), a glucose molecule labelled with radioactive fluorine (F-18). The PET images used in this paper were also obtained by administering FDG and capturing the resulting images. This PET imaging technology enables the examination of blood flow, metabolism, and neural transmission in organs such as the brain and liver, making it the subject of various studies [28][29][30]. Physicians construct three-dimensional data in their minds from approximately 300 PET cross-sectional images of the entire body and diagnose by identifying organs based on the intensity and characteristics of accumulations (clusters of high-intensity points). In PET imaging, organ segmentation is crucial as it enables examinations considering the physiological metabolic characteristics of each organ. When interpreting PET images, physicians first identify the organ and then determine whether it is normal or abnormal. Figure 3.2 shows examples of PET cross-sectional images. These are PET cross-sectional images including the kidneys, liver, and lungs. As shown in Figure 4, PET images have blurred boundaries, making it difficult even for the human eye to identify the regions of each organ without specialised knowledge or experience. However, when considering a PET image diagnosis support system, organ segmentation is an indispensable image processing step. In this section, we focus on the regions of the kidneys, lungs, and liver in PET images. The kidneys are two strong accumulations in the lower abdomen, the lungs are two weak accumulations flanking the heart, and the liver is a moderately strong accumulation between the kidneys and the heart. Each has distinctive physiological characteristics in terms of accumulation intensity, but as shown in Figure 4F, it is impossible to identify the regions using simple binary segmentation based on grayscale values. Additionally, the liver resembles the heart in terms of intensity and characteristics, requiring specialised knowledge of location and shape for differentiation. This segmentation processing of the kidneys, lungs, and liver will be constructed using 3D-ACTIT.

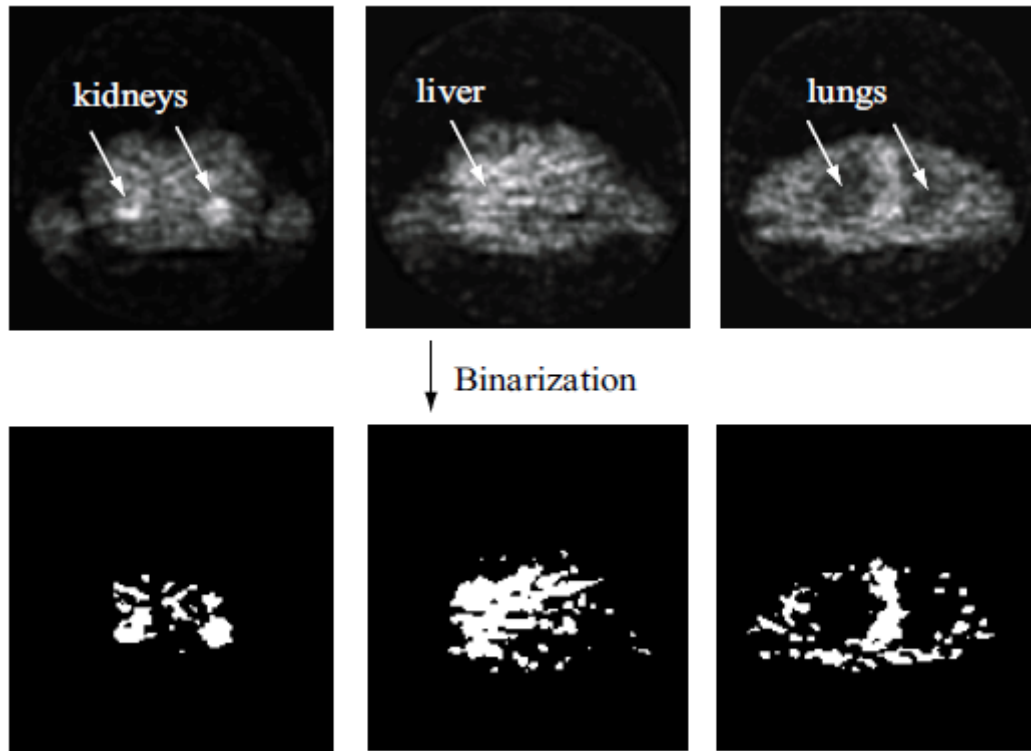


Figure 4 Examples of PET image

Diffusion-weighted image

Diffusion-weighted images (DWI) are images that visualise the diffusion motion of water molecules that constitute living organisms. As fast imaging techniques for MRI, the fast spin echo method and the fast gradient echo method are representative, with the former being widely used for fast imaging of T2-weighted images and the latter for MR angiography. On the other hand, echo-planar imaging enables the acquisition of images similar to those obtained using the spin echo method or gradient echo method in 40–50 milliseconds per slice. Diffusion-weighted images are sensitive to motion and are highly prone to motion artifacts. Therefore, performing diffusion-weighted imaging using echo-planar imaging allows the acquisition of diffusion-weighted images that are less affected by motion. Diffusion-weighted images have a completely different physical basis from T1- and T2-weighted images [31]. By using this diffusion-weighted imaging technique, it has become possible to detect conditions such as cerebral infarction and cytotoxic oedema, which were previously difficult to visualise with different contrast using conventional methods. Currently, extensive research is being conducted in the medical community [32], contributing new insights to the imaging diagnosis of conditions such as haemorrhage, demyelination, tumour lesions, infections, trauma, and epilepsy.

Figure 5 shows an example of diffusion-weighted images. The images used in this paper were acquired using a GE iMRI scanner with the following parameters: TE: 120 msec, TR: 5000 msec, Bandwidth $\pm$: 62.5 kHz, and bvalue: 1000. As shown in Figure 5, abnormal signals are clearly visualised. However, while this technique effectively visualises abnormal signals compared to

T1- and T2-weighted images, not all high-intensity regions are necessarily abnormal. Figure 5(a) shows an example from a healthy individual, and (b) shows an example containing abnormal signals. Both images include high-intensity regions. When diagnosing these cases, physicians determine whether the high-signal regions are magnetic susceptibility artefacts or abnormalities based on brain symmetry, knowledge, and experience. In this study, we aim to extract only abnormal signals from diffusion weighted images and present the results of our experiments.

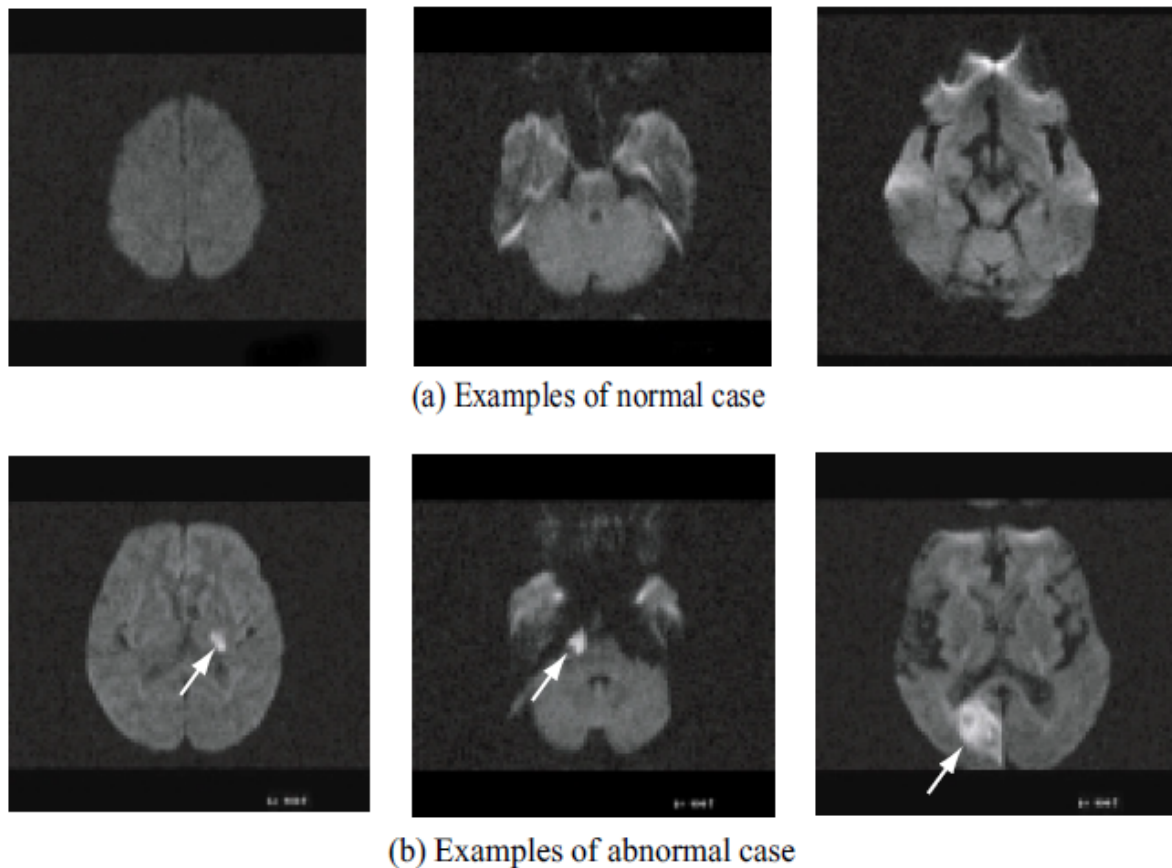


Figure 5 Examples of diffusion-weighted image

2.2.6. Experiments on organ segmentation from PET images

A set of training images was created with the aim of constructing segmentation processing for the kidneys, lungs, and liver from PET images, and each was input into 3D-ACTIT. The training image set for kidney and lung region extraction is shown in Figure 6. The training image set consists of 100 PET cross-sectional images with a resolution of 128×128 pixels, spanning from the shoulders to the waist, forming a 3D image dataset. This training image set includes regions such as the kidneys and mediastinum, which show strong accumulation, the liver and heart, which have weaker accumulation but are larger than the kidneys, and the lungs, which show very weak accumulation. As shown in Figure 6, the target images for kidney region extraction were created using binary images where only the kidney region was coloured white

and the rest was black. Similarly, binary images were used to create target images for the lung region. These target images were created with the assistance of a physician. The parameters of the GP are shown in Table 3. The termination conditions used in the experiments in this section were set to either when the fitness reached 1.0 or after 3,000 generations, which was confirmed to be sufficient convergence based on previous experiments.

Table 3 GP parameters in 3D-ACTIT

Parameter	Value
The number of generations	3000
Population size	30
Crossover rate	1.0
Mutation rate	0.9
Selection	Tournament selection
Tournament size	2

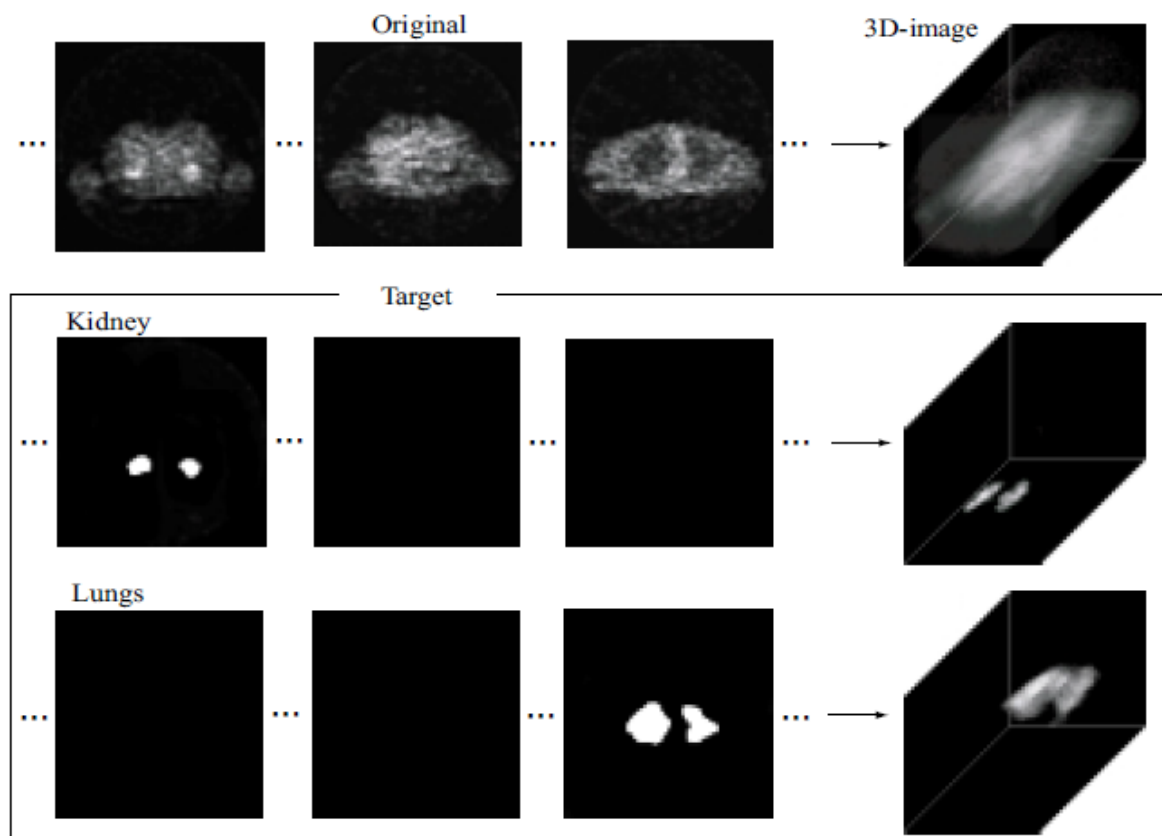


Figure 6 Training data set for extracting kidney and lung regions

Kidney region extraction from PET image

A training image set targeting the kidney region shown in Figure 6 was input into 3D-ACTIT. The output images from the training image set are shown in Figure 7, and the constructed tree-structured filter and intermediate results of the image conversion are shown in Figure 8. In Figure 8, the original images are input from all leaf nodes of the tree structure and processed by each image processing filter from the lower layer to the upper layer, ultimately outputting a single image. Analysing the tree structure, in the lower layer of the tree structure, as shown in Figure 8(a), the average value filter "-" is applied to the intense product "u" of the original images, simultaneously emphasising accumulations such as the mediastinum and liver. On the other branch, the same intense product of original images is applied to a minimum value filter, as shown in Figure 8 (b), emphasising only the strongest clusters of accumulation, even if they contain noise. By taking the limit product "b", as shown in Figure 8 (c), it can be seen that the kidney region is roughly extracted, with other organ accumulations and noise removed. In the latter half of the processing, as shown in Figure 8 from (d) to (f), steps such as neighbourhood processing using maximum value "M", minimum value "m", and average value filter "-", as well as 2-input filters to refine the shape, are observed. Finally, a binarization filter is applied to output the image, bringing it closer to the target image. When comparing the target image and the output image, it can be seen that the results are very close, with an adaptability (fitness) of 0.999 after 3,000 generations. Regarding processing time, several days are required to optimise the processing procedure. However, once the filters are constructed, the process can be executed in approximately 3 seconds using a general-purpose computer for the image data size and filter size used in this study.

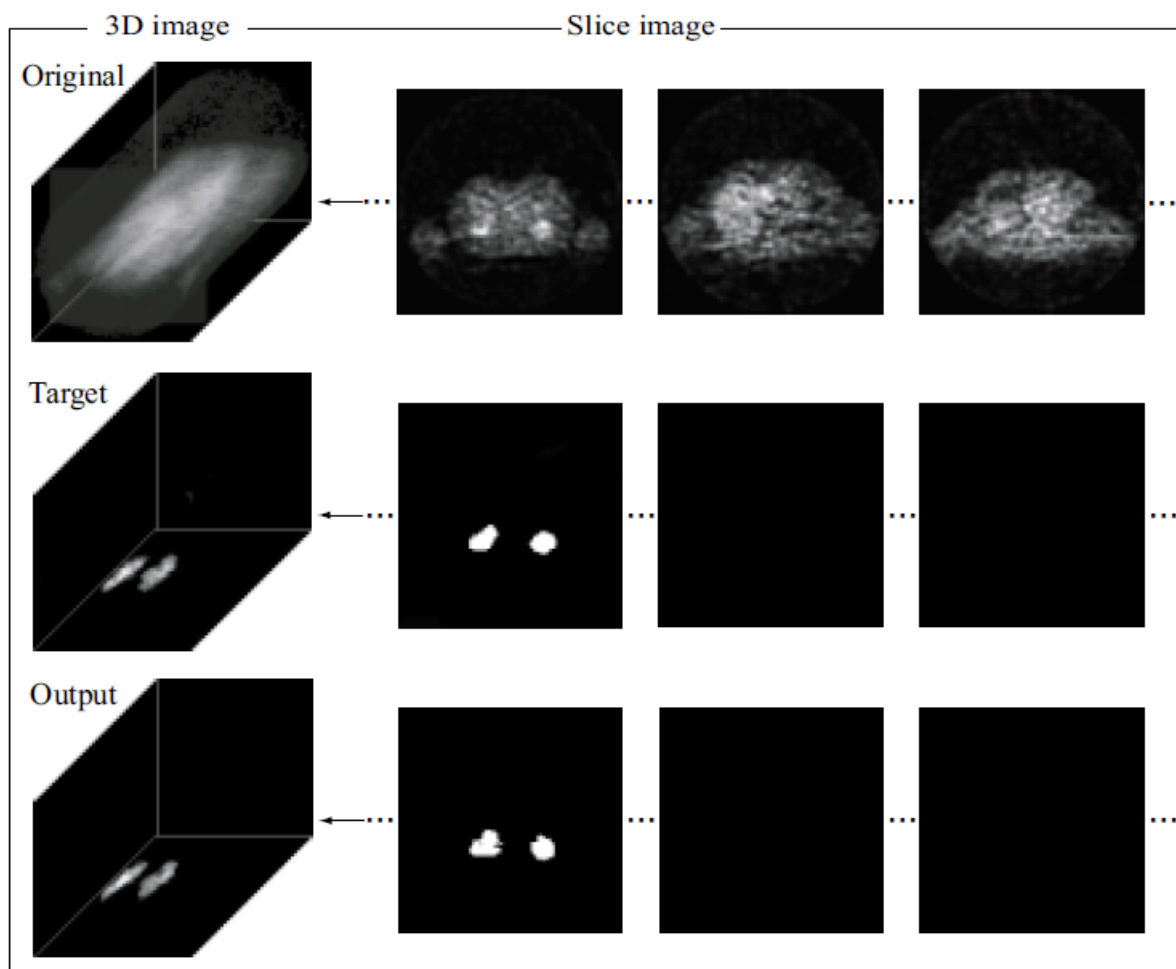


Figure 7 Results on training data for kidney region extraction

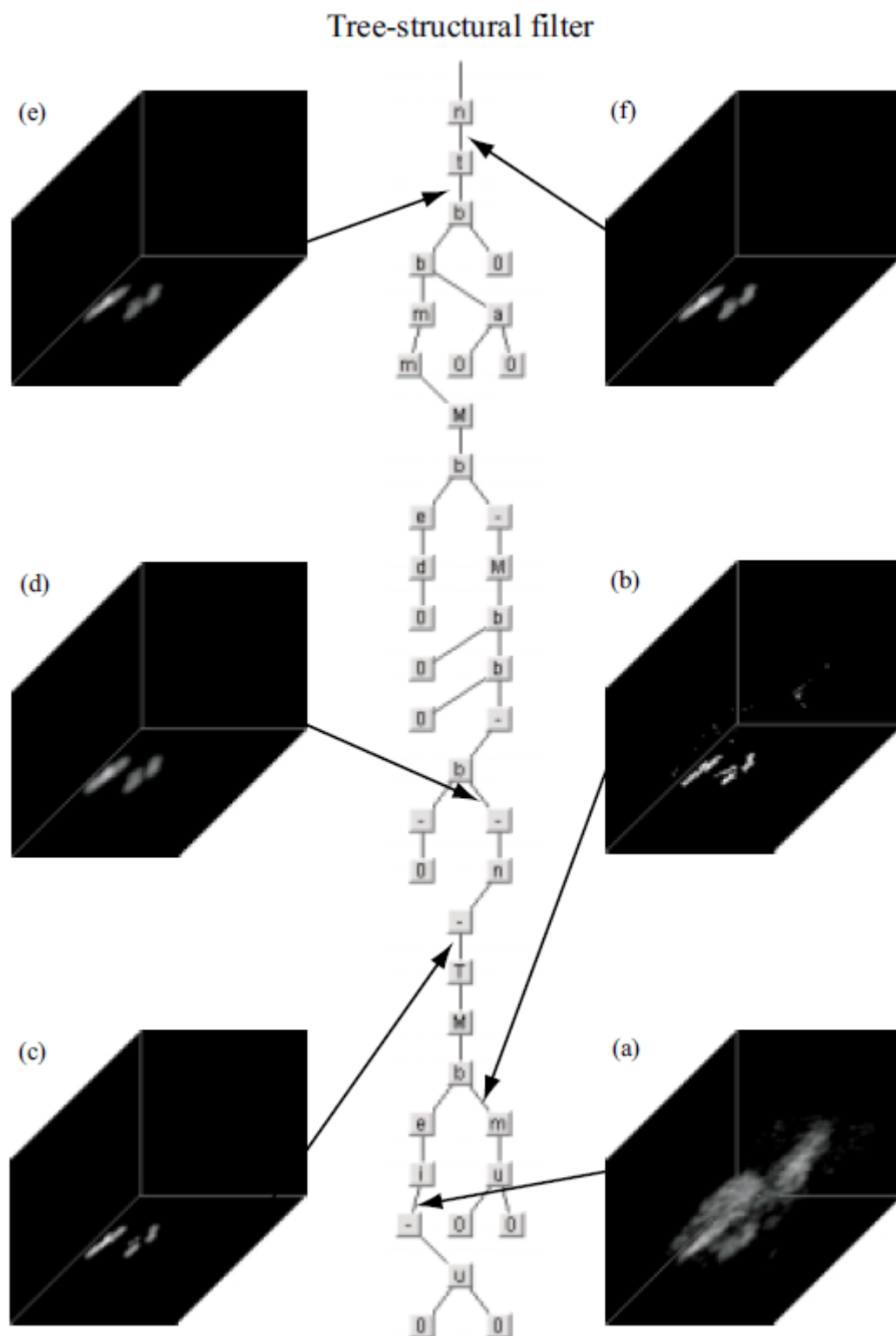


Figure 8 Obtained tree-structural filter in kidney extraction experiment

Lung region extraction from PET image

A set of training images targeting the lung region shown in Figure 6 was input into 3D-ACTIT. The output images from the training image set are shown in Figure 9, and the constructed tree-structured filter and intermediate results of image conversion are shown in Figure 10. In this experiment as well, after 3,000 generations of learning, the fitness reached a high value of 0.991. The constructed tree-structured filter took on a linear shape centred around a 1-input 1-output filter. Analyzing the tree structure from the bottom up, the image is first smoothed using an average filter "-" and then binarized to create a mask image targeting only the human body, as shown in Figure 10 (a). By taking the algebraic product "a" of this masked image and the image in Figure 10 (b), which emphasises areas with fewer accumulations inside the body using a white edge enhancement filter "E", we obtain an image in Figure 10 (c), which contains noise but roughly extracts the lung region. In the upper part of the tree structure, as shown in Figure 10(d) and (e), noise is removed by combining neighbourhood processing filters such as the average filter "-" and the minimum filter "m". As shown in the output image in Figure 9, we successfully extracted only the lung region. However, when comparing the output image with the target image at the cross-sectional level, it can be seen that the shape is slightly distorted. This is because the minimum value filter, which is used to remove noise, also affected the lung region that had already been extracted.

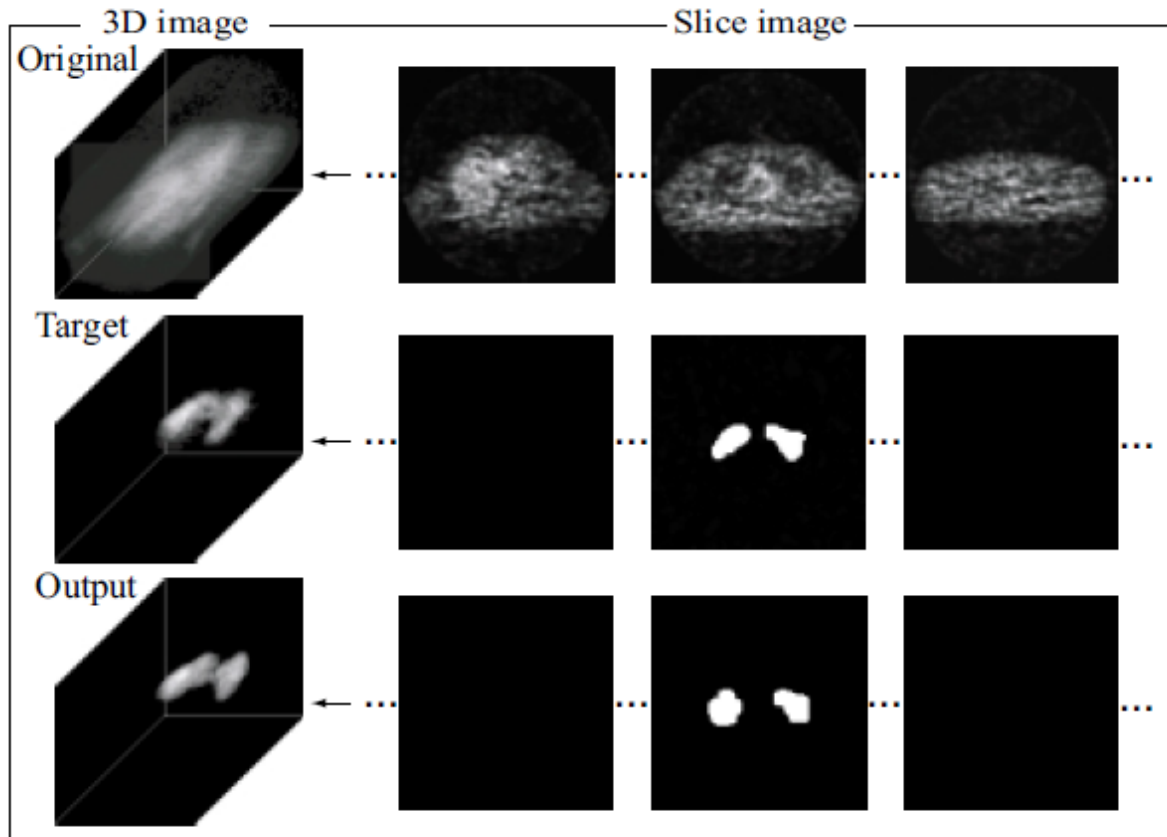


Figure 9 Results on training data for lung region extraction

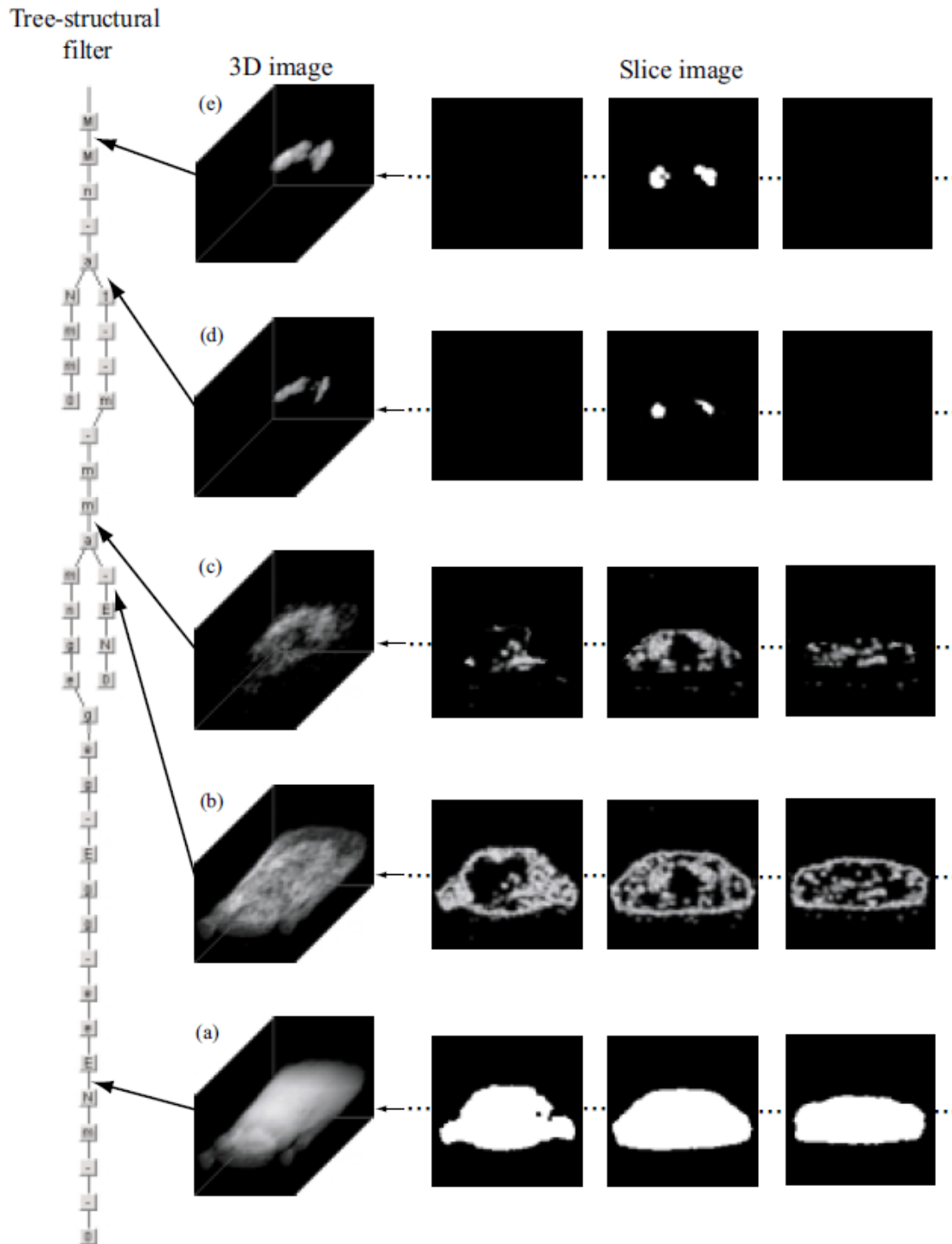


Figure 10 Obtained tree-structural filter in lung extraction experiment

Application to testing data

There are individual differences in the position and size of organs in the human body, and especially in functional images such as PET images, the signal strength may vary slightly depending on the health condition of the same individual at the time. Therefore, it is

necessary to verify whether the constructed tree-structured filter can handle signal strength variations caused by such individual differences or health conditions. In experiments to extract kidney and lung regions, the tree-structured filters constructed for each region were applied to images not used for training to verify the robustness of the tree-structured filter. To evaluate the results and compare them with correct examples, we also wanted to confirm the shape of the extracted regions and the presence or absence of noise, which are difficult to represent using the fitness function in Equation (2). Therefore, we conducted a visual evaluation in collaboration with the physicians who provided the images. Several examples of the results of applying the filters to images not used for training are shown in Figure 11. As can be seen from Figure 11, good extraction results were obtained, particularly for the kidneys. In the lung region extraction experiment, we were able to extract only the lung region without reacting to other organs or accumulations, however there were two cases where insufficient extraction of the left lung was observed, as shown in Figure 11(a-3) and (d-3). This was because, in the tree structure shown in Figure 10, during the latter half of the processing, from Figure 10(c)(d), noise was partially removed as noise during the conversion process. Since the heart is located more to the left than the centre, the left lung is smaller in volume than the right lung. The fact that the target area may be mistaken for noise during the processing process is a drawback of this tree structure. In this experiment, only one example was used as the training image set, therefore there were cases where individual differences in organ size and signal strength could not be accommodated. By using multiple examples as the training image set, it is considered possible to construct a more robust algorithm for testing images.

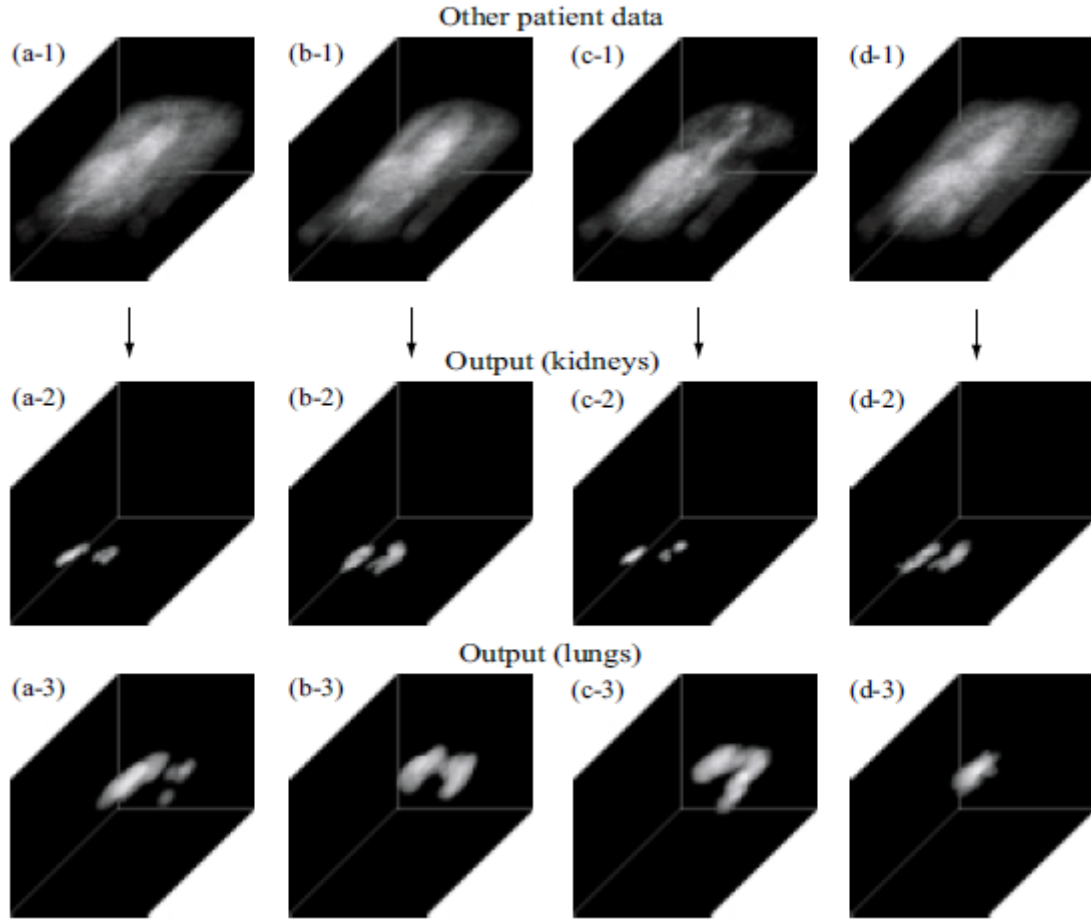


Figure 11 Results on testing data

2.2.7. Experiments by using 3D-ACTIT with filters using target regional features

In the previous sections, we demonstrated the basic performance of 3D-ACTIT through experiments. However, since the basic filter set of 3D-ACTIT primarily consists of neighbourhood processing and pixel value mapping, while it can extract features with distinctive pixel values, such as kidneys and lungs in PET images, it becomes challenging to distinguish the target object from other objects with similar pixel values or sizes within the image space. In this section, we added filters using candidate region features to the basic filter set of 3D-ACTIT and conducted experiments. The candidate region features used here are calculated through labelling processing, such as position and volume in 3D space. Although these are simple features, the position and size of organs within the human body vary slightly among individuals but are not significantly different. In the experiments conducted in this section, liver extraction from PET images and abnormal signal extraction from diffusion-weighted images, image processing filters that can consider the position and symmetry of organs were added to the filter set, then, experiments were conducted. The combination of the basic filter set's neighbourhood processing and pixel value mapping filters with the newly

added filters demonstrates the construction of more advanced and versatile image processing capable of handling testing images.

Liver region extraction from PET image

A set of training images was created for the purpose of extracting the liver region from PET images. Figure 12 shows the training image set. As can be seen from Figure 12, the voxel value characteristics of the liver and heart are very similar, and the size difference in three-dimensional space is also small, requiring a complex algorithm to extract the liver alone. Additionally, the liver exhibits stronger aggregation than other organs and tissues, making it difficult to distinguish from the kidneys. Therefore, in this experiment, we prepared a new filter using candidate region features as shown in Table 4. We created an image processing filter that uses information such as the liver's position on the right side of the body and its larger size compared to the kidneys as features for object position and size. Furthermore, since the liver has a highly distinctive shape, edge enhancement filters for each direction were prepared to improve extraction accuracy. These filters were added to the basic filter set from the previous section, and experiments were conducted to determine the optimal combination, which was ultimately decided as 3D-ACTIT.

Table 4 Additional 3D image processing filters for liver segmentation

Image Processing Filters Using Candidate Region Features	
f	Remove objects located on the left half of the centroid
P	Remove objects located on the right half of the centroid
r	Remove objects with volume smaller than or equal to the threshold
R	Remove objects with volume larger than or equal to the threshold
Directional Edge Enhancement Filters	
x	Sobel filter in the x-direction
y	Sobel filter in the y-direction
z	Sobel filter in the z-direction

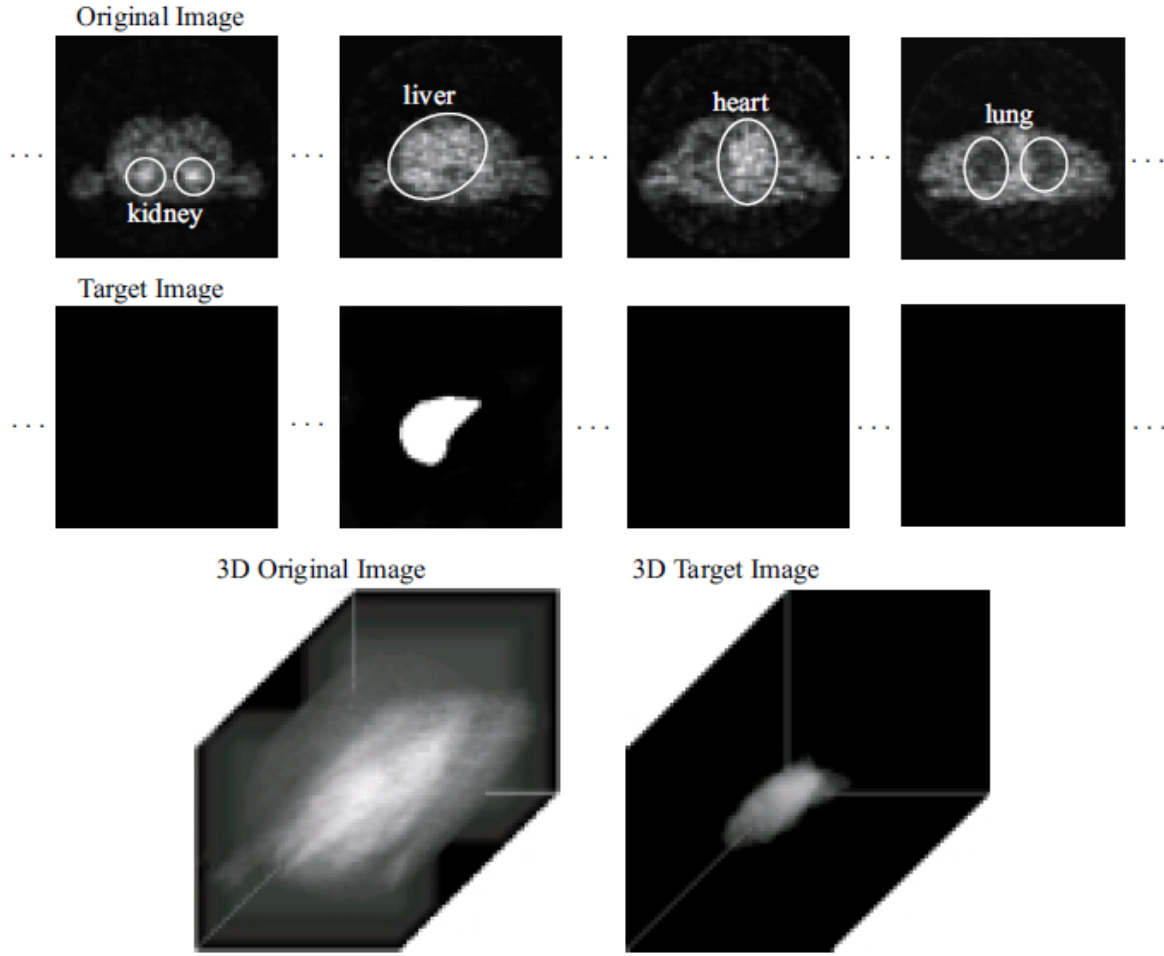


Figure 12 Training data set for extracting liver regions

The training image set in Figure 12 was input into 3D-ACTIT, and the resulting output images are shown in Figure 13. The constructed tree-structured filter and intermediate results of image conversion are shown in Figure 14. This tree-structured filter was constructed through 3,000 generations of learning, with an adaptability of 0.995. The final output image in Figure 13 is very close to the target image. By examining the input images at the ends of the tree in order, we can see that the liver region and heart region with similar aggregation features are first extracted using neighbourhood processing such as the minimum value filter "m" and logical product "l" or algebraic sum "A". Next, the image is shaped using filters such as the Sobel filter "d" and the average filter "-", then binarized "n", and finally, the filter "f" utilising the positions of the clusters is selected to successfully retain only the liver region. As shown in the cross-sectional image of the output results in the same figure, although there is some deformation, the characteristic shape of the liver is also successfully captured.

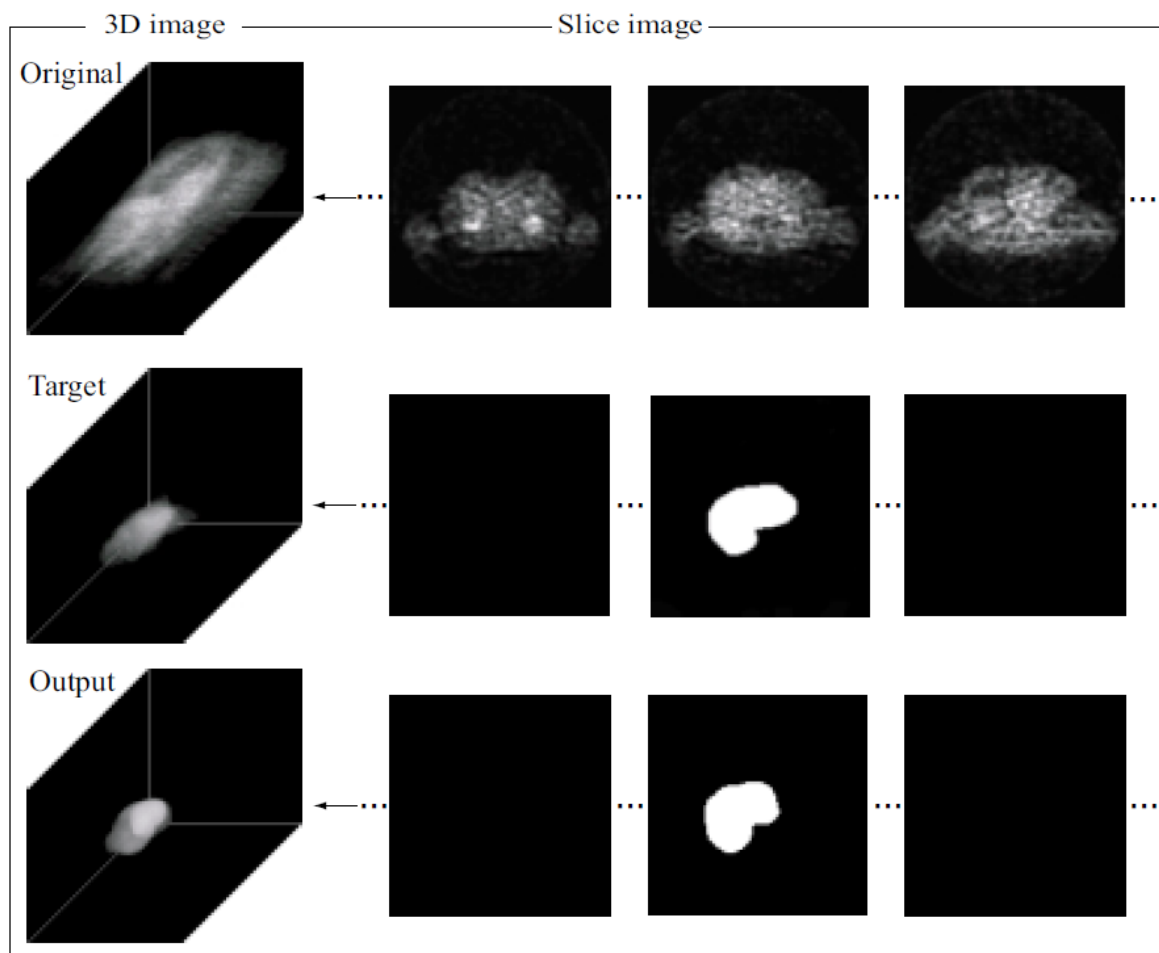


Figure 13 Results on training data for liver region extraction

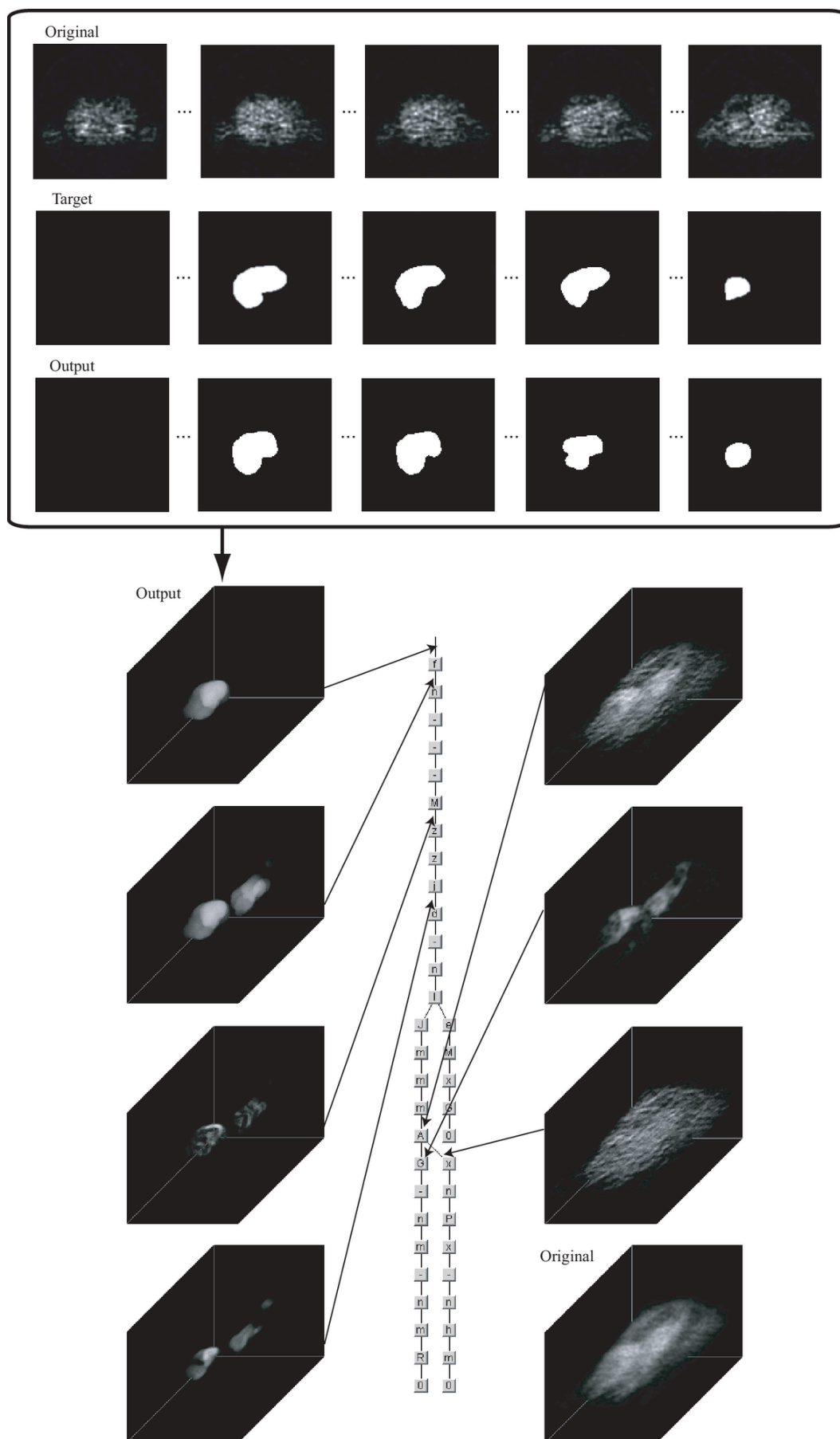


Figure 14 Obtained tree-structural filter in liver extraction experiment

Figure 15 shows the transient in fitness and output images before and after adding filters that consider candidate region features in the liver region extraction experiment. In both experiments, fitness increases as learning progresses. However, when comparing the results before and after adding filters that utilise features, it can be seen that better results are obtained by utilising features even with the same training image set. Before adding filters that consider features, the pixel value conversion filter was the main filter, therefore organs with similar accumulation features could not be distinguished. When combining filters to extract only the liver, removing accumulations other than the liver using a uniform pixel value conversion process across the entire image caused the liver to also be affected by the process. However, by using candidate region features, it is possible to determine whether an object is the target or not based on information other than grayscale values, such as position and size. This prevents deformation of the target organ, as seen in previous experiments, and is believed to have led to the results obtained. This trend in fitness was observed in multiple experiments using the same training image set and in experiments using other patient data. By using filters that consider features, it was possible to extract the characteristic shape of the liver and achieve high fitness.

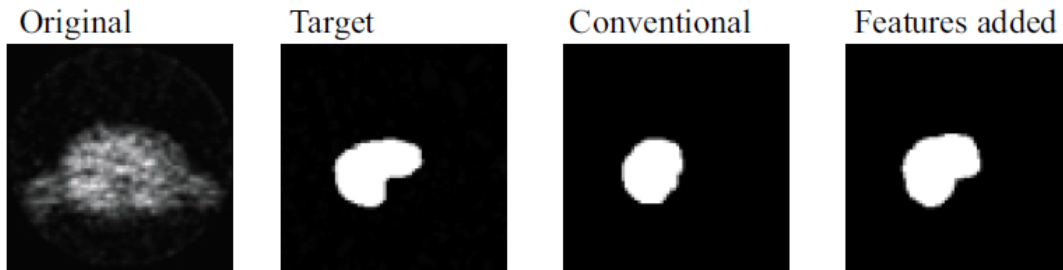
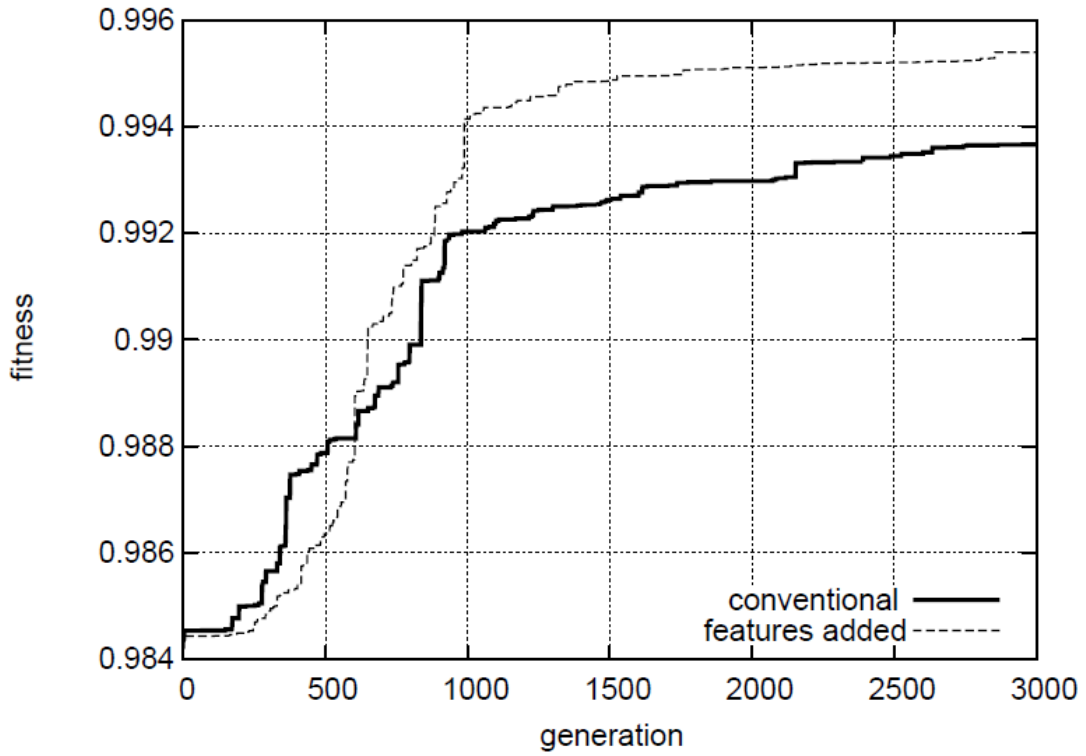
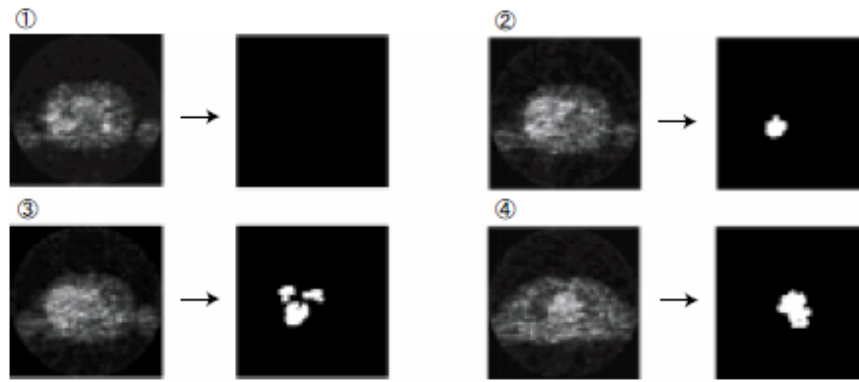
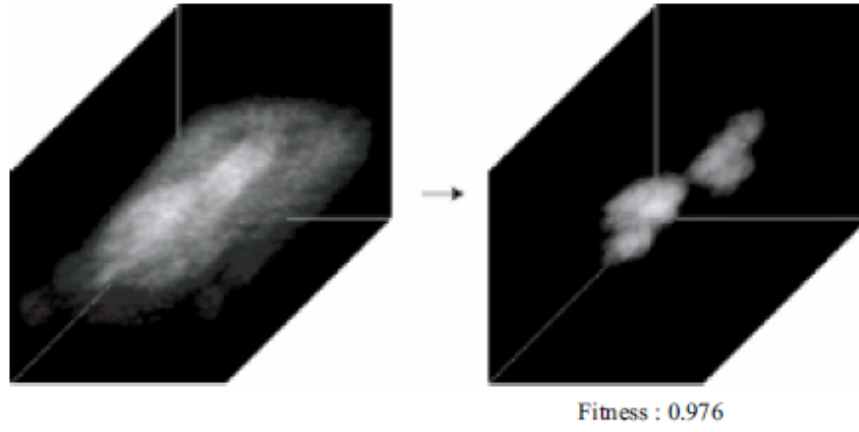


Figure 15 Transients of fitness with and without liver feature

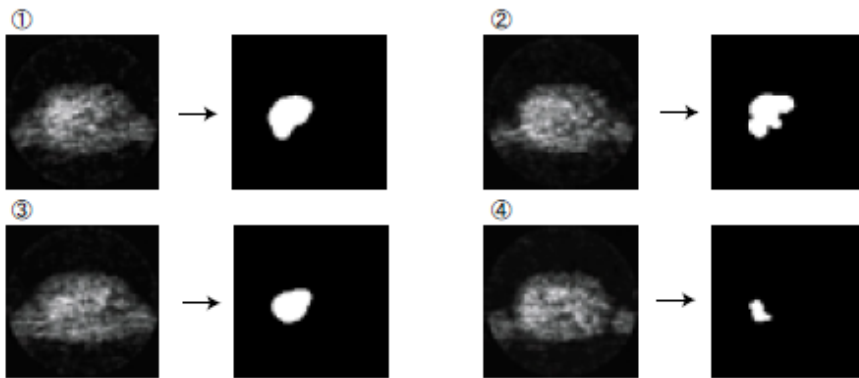
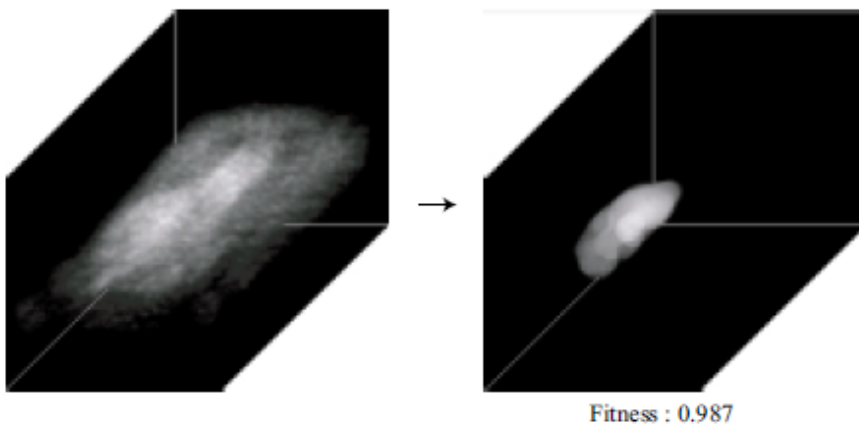
Application to testing data

Results of applying the constructed tree-structured filter to image data from other patients not used for learning are shown in the Figure 16 shows the results. Figure 16(a) shows the results obtained by applying the filter obtained from 3D-ACTIT before adding filters using features, and Figure 16(b) shows the results after adding filters using features. Before adding filters using features, accumulations other than the liver were also extracted, however after adding filters using features, only the liver area could be extracted even for images that had not been learned. When comparing the cross-sectional images, before applying the feature-based filter addition, the shape of the liver is distorted, and the area where the liver first appears when viewed from the bottom to the top of the body is not extracted. On the other hand, after adding filters using features, the liver was extracted from the beginning to the end, and no other organs were over-extracted. Furthermore, when comparing the fitness, the output before adding filters using features had a fitness of 0.976, while the output after adding filters using features had a fitness of 0.987. This difference of 0.011 in fitness is due to the

large image size of image size is $128 \times 128 \times 100$, which is large, therefore, although the difference is small in numerical terms, it is significant in the image. Furthermore, when the experimental results, including these, were evaluated by a physician, they were judged to be "extremely effective." In terms of visual shape evaluation, quantitative evaluation based on fitness, and evaluation by a physician, the system with added features yielded better results in all cases.



(a) Conventional 3D-ACTIT



(b) 3D-ACTIT with regional features

Figure 16 Results on testing data for liver region extraction

Diffusion-weighted image processing by 3D-ACTIT

Considering candidate region features in diffusion-weighted imaging experiments. First, Figure 17 shows examples of diffusion-weighted images containing high-signal regions caused by magnetic susceptibility artefacts and examples containing abnormal high signals. Determining whether each signal in diffusion-weighted images is a magnetic susceptibility artefact or an abnormal signal is extremely important. When considering the construction of a system to support physicians, it is essential to extract each signal and recognise only abnormal signals. As shown in Figure 17, the brain generally exhibits symmetry in both high-signal regions caused by magnetic susceptibility artefacts and other signals. However, the location of abnormal signals is not fixed. In this experiment, we focused on the symmetry of signals in the brain and added an image processing filter using candidate region features to 3D-ACTIT to extract only abnormal signals. Table 5 shows the added image processing filters. These filters use the 3D spatial position and volume of signals as candidate region features and apply processing to retain or remove symmetrical objects based on the centre of mass coordinates of the entire brain calculated from the original image.

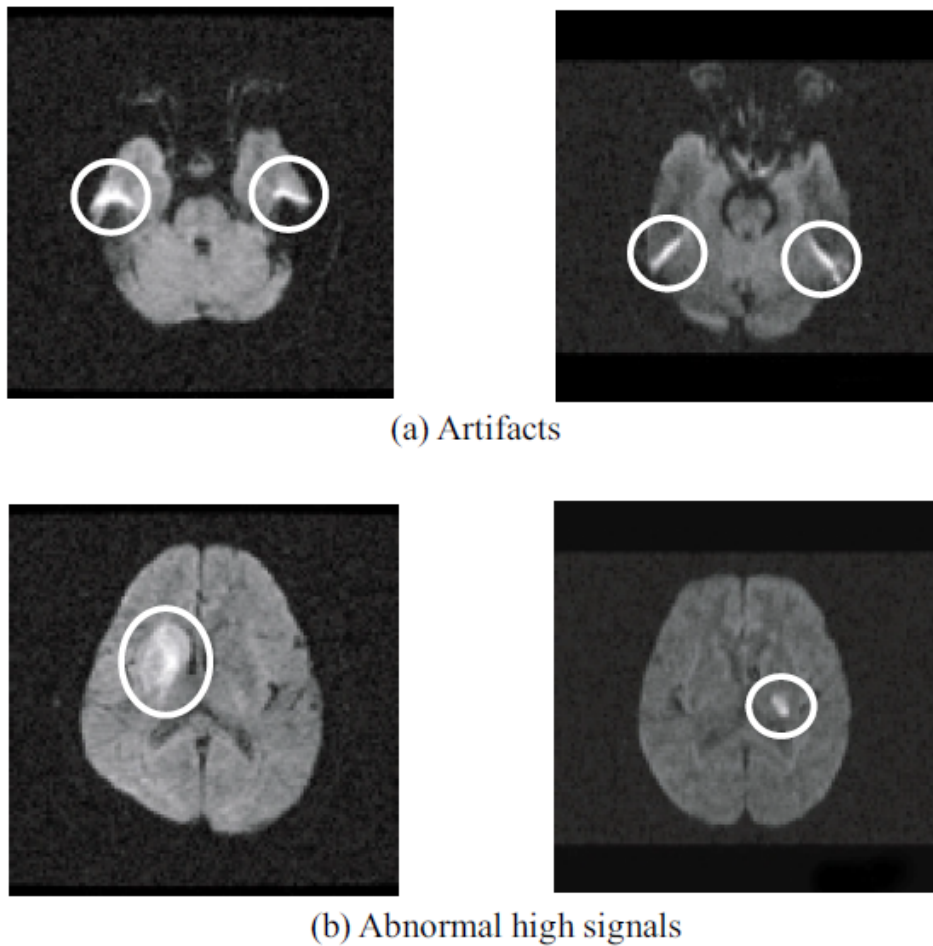
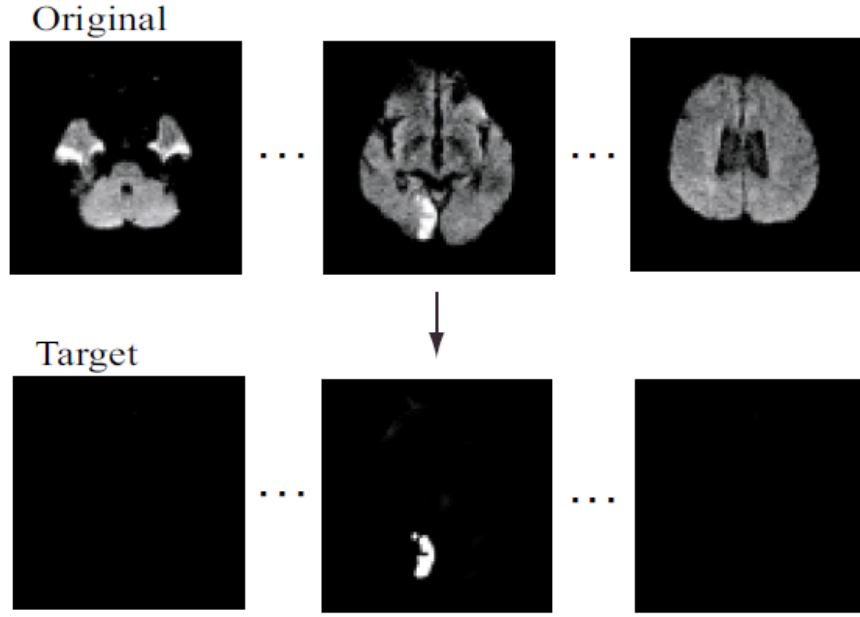


Figure 17 Artifact and abnormal signals in diffusion-weighted image

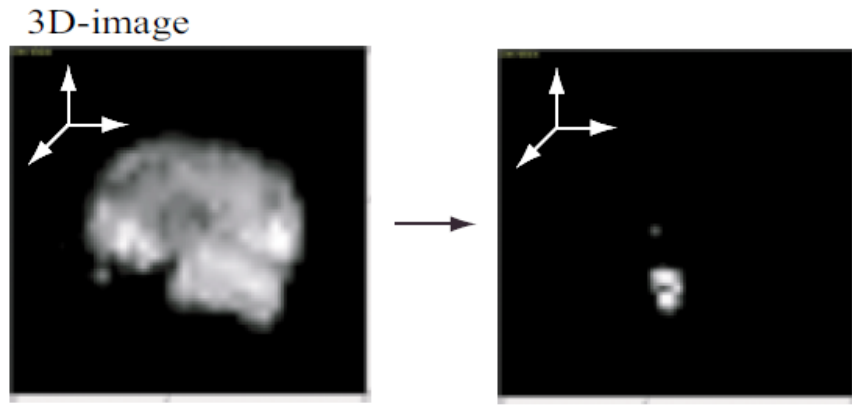
Table 5 Additional 3D image processing filters for asymmetry signal extraction

Image Processing Filters Considering Candidate Region Features	
k	Remove symmetric objects
K	Remove all except symmetric objects

A set of training images was created with the aim of extracting abnormal signal regions from diffusion-weighted images, and these were input into 3D-ACTIT. An example of the created training image set is shown in Figure 18. The signal to be extracted was coloured white, and the rest was black using a painting tool. As shown in Figure 18, the cross-sectional view is easier to interpret, therefore, although the processing is performed in three dimensions, all images will be displayed as cross-sectional views in subsequent experiments. This training image set was created using diffusion-weighted images obtained from 18 cross-sectional views of a single patient.



(a) Slice images



(b) 3D voxel data

Figure 18 Training dataset for extracting abnormal signals from diffusion-weighted image.

The results obtained by inputting the training image set in Figure 18 into both 3D-ACTIT before and after adding the filters using candidate region features are shown in Figure 19. As a result, the output image converted from the original image and a tree-structured filter representing the image conversion algorithm, as shown in right hand side of Figure 19, were obtained. 3000 Generation calculations were performed, and the adaptability of the output images obtained was high at 0.99 for both. As can be seen in Figure 19, we were able to obtain 3D image conversion that produced output results very close to the target image both before and after filter addition. Although not all cross-sectional images are shown in Figure 19, no noise other than the target signal was observed in the other cross-sectional images.

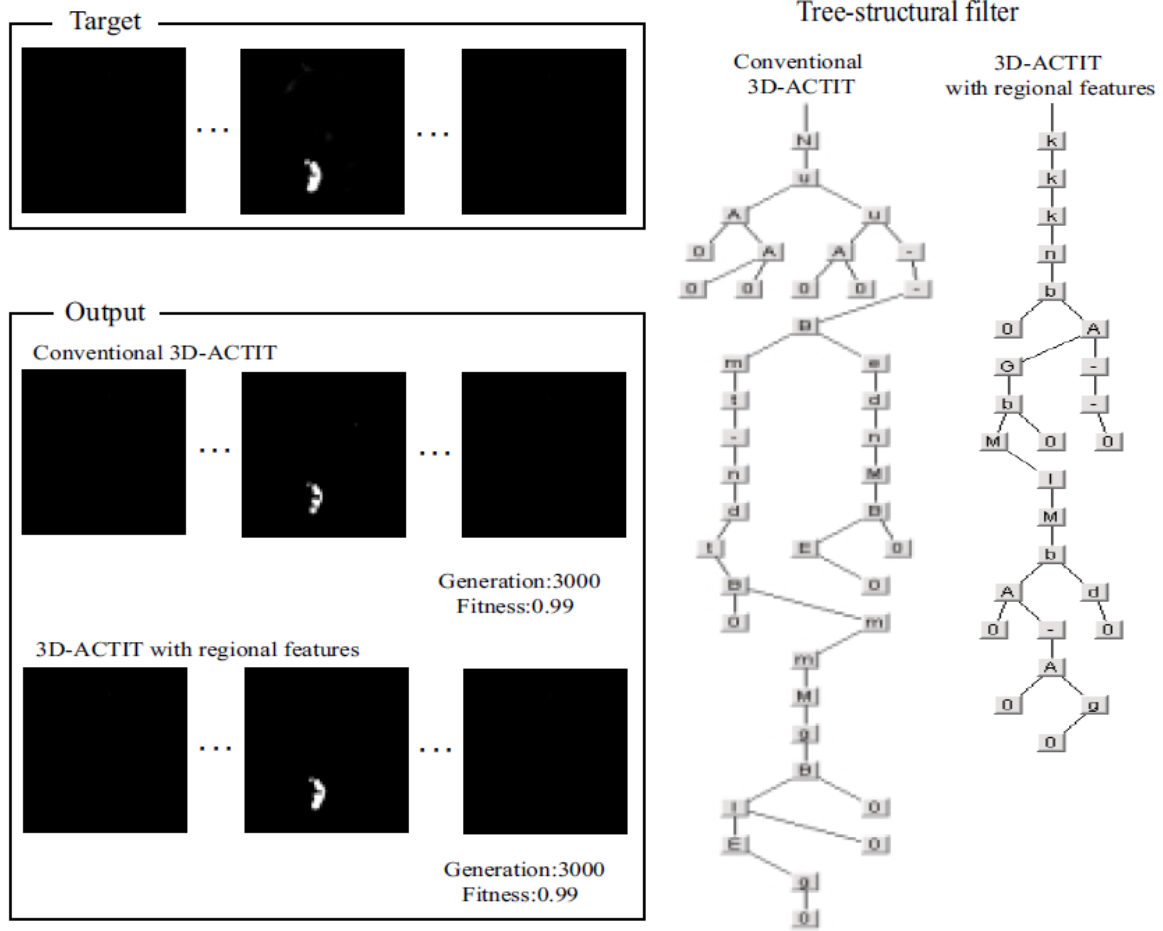


Figure 19 Results on training data for abnormal signal extraction on diffusion-weighted image

The results of applying the tree-structured filter obtained in Figure 19 to testing images not included in the training image set are shown in Figure 20. The results are shown in Figure 20. Cross-sectional images containing abnormal signals and high-signal regions other than abnormal signals are selected and marked. Testing images are also 3D image data consisting of 18–20 cross-sectional images, similar to the training images. Looking at the output results before adding candidate region features, images where abnormal signals can be extracted also extract high-signal regions caused by magnetic susceptibility artefacts. Additionally, in cases where the abnormal signal region is extremely large, extraction was not possible. The basic filter set of 3D-ACTIT primarily consists of pixel value conversion processing filters, and for the few examples used as the training image set, the order of the pixel value conversion processing filters was trial-and-errored to construct the desired processing. However, when applying this to testing images, there are cases where the expected processing cannot be performed due to significant differences in overall image contrast or differences in the shape and position of the objects to be extracted. In other words, using a combination of filters centred on pixel value conversion processing may result in the construction of processing specific to the training images. On the other hand, after adding candidate region features, we were able to extract abnormal signals with different sizes, shapes, and positions without

omission. Additionally, we could extract only abnormal signals without over-extracting high-signal regions caused by magnetic susceptibility artefacts. While the basic filter set struggled to adapt to changes in contrast or object features, adding candidate region features enabled the construction of a more general-purpose processing method capable of addressing such variations.

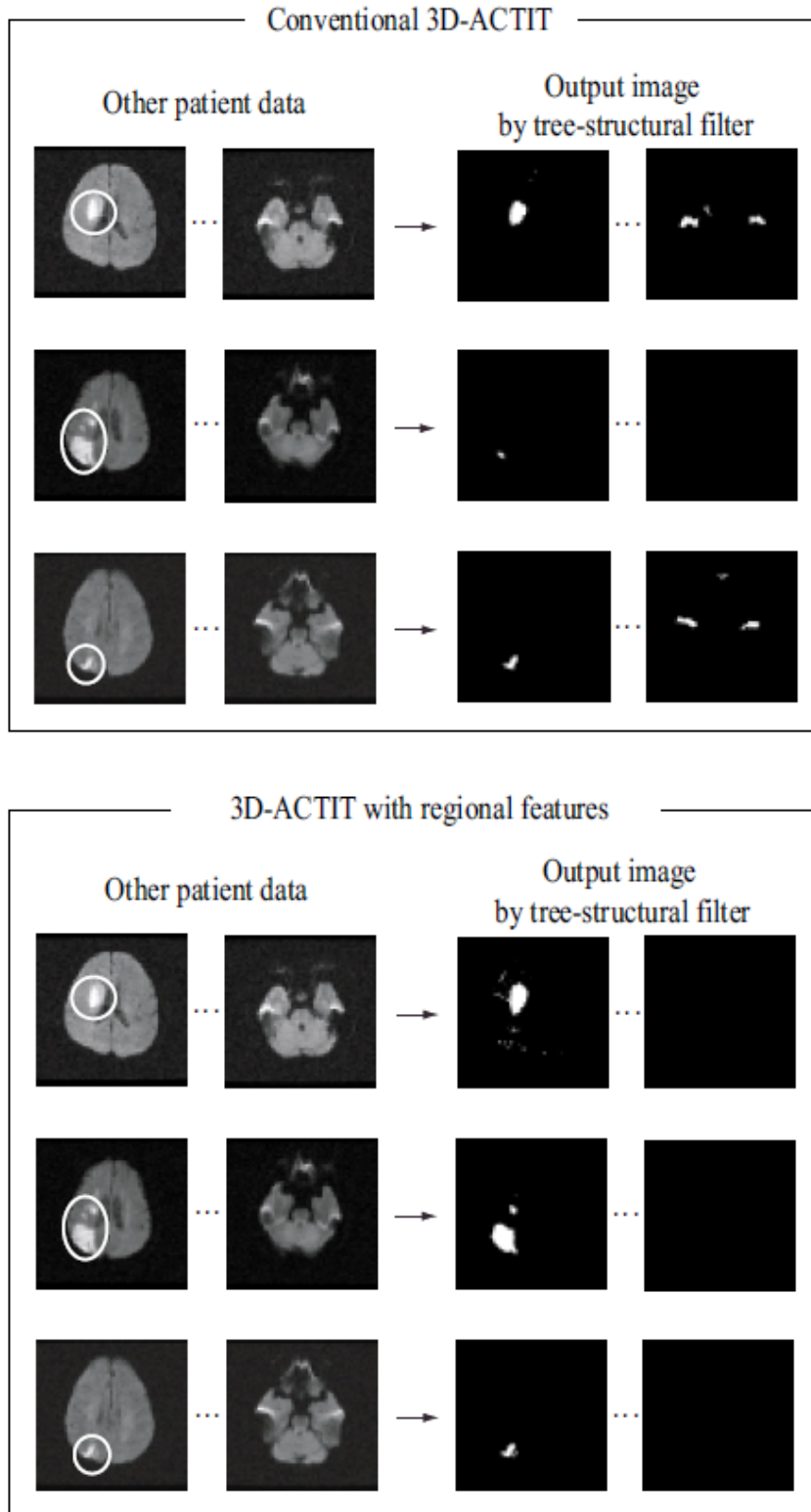


Figure 20 Results on testing data

Figure 21 shows the detailed analysis results of the tree-structured filter constructed using 3D-ACTIT after adding filters considering candidate region features. The original image "0" is input from the end of the tree structure, and image conversion is performed by three-dimensional image processing filters assigned symbols, resulting in a single output. To analyse

the results in detail, we examine the processing results in order from the bottom-most input image in the tree structure. First, in the early stages of processing, edge extraction processing such as the Laplacian filter "g" and Sobel filter "d", and neighbourhood processing such as the average filter "-" and maximum filter "M" are performed.

"A" to widen the brightness difference between high-signal areas and areas with relatively weak signals. The images with this brightness difference are then used as a mask and passed through an intense product filter to extract the high-signal areas. At this stage, both abnormal signals and other high-intensity regions are still extracted. However, in the latter half of the processing, a filter that considers candidate region features—specifically, the "k" filter, which eliminates objects with symmetrical positions and volumes—is selected to retain only the abnormal signal regions.

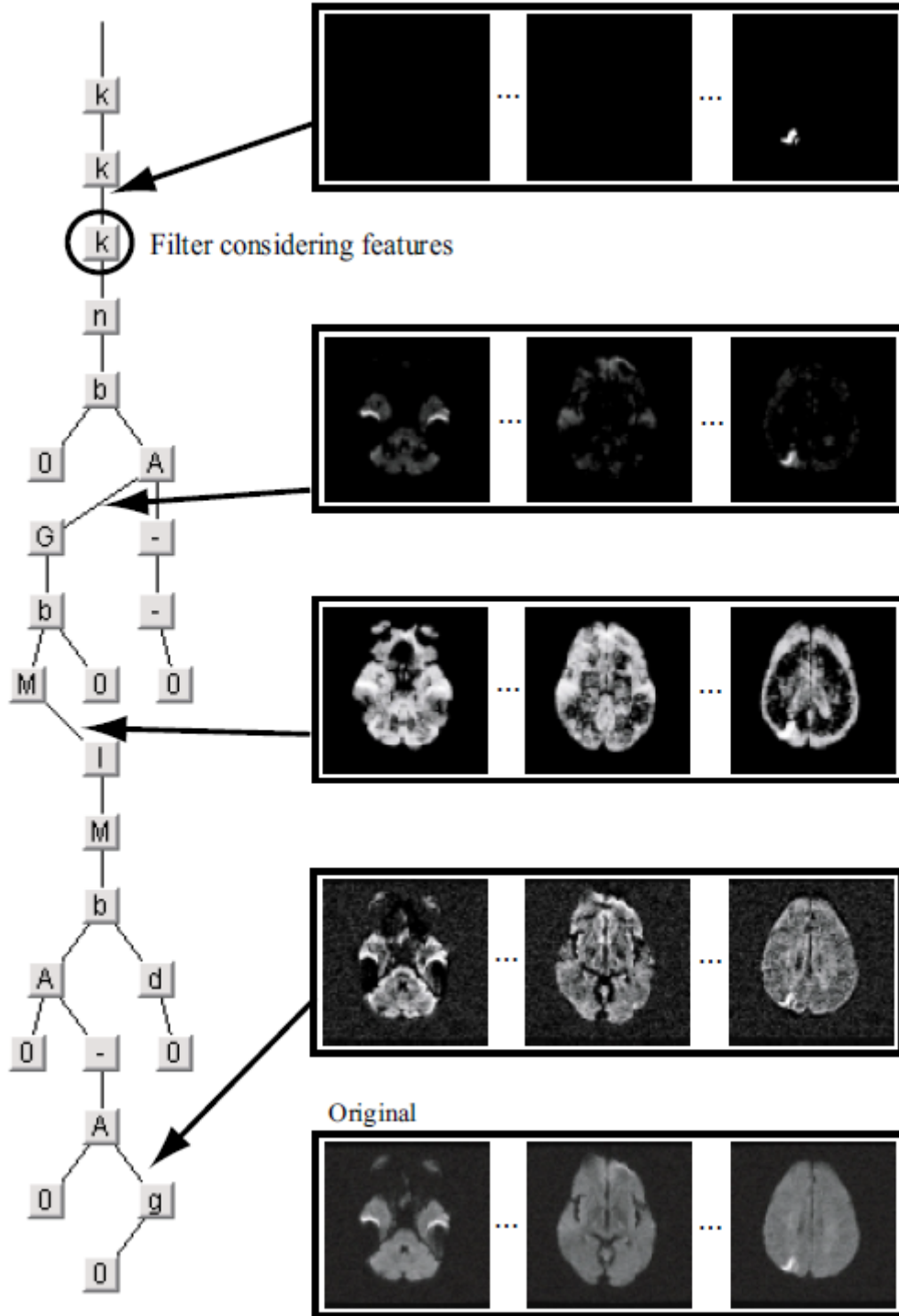


Figure 21 Obtained tree-structural filter in liver extraction experiment

2.2.8. Summary

In this section, we proposed 3D-ACTIT, which automatically acquires complex and effective 3D image processing that cannot be achieved by simply arranging basic 3D image processing filters in a tree structure using GP. Through experiments using PET and diffusion-weighted

images, we evaluated the effectiveness of the system by collaborating with physicians to assess the results of applying the constructed tree-structured filters and images not used for learning. Furthermore, by adding image processing filters that utilise image features, we were able to efficiently separate organs with similar accumulation characteristics, such as the liver, in PET image processing. In diffusion-weighted image processing, we successfully extracted only abnormal signals, demonstrating its effectiveness. We were also able to perform the expected processing on images not used for learning, confirming the robustness of the constructed tree-structured filters. Furthermore, 3D-ACTIT was applied to dynamic image processing, suggesting its potential (Appendix A).

In the future, we aim to enhance the image processing filter set by adding filters that utilise grayscale conversion and image features, and to develop more advanced image processing algorithms. Additionally, we plan to improve the 3D-ACTIT system to incorporate medical or image-related expertise when handling medical images like those used in this study.

2.3. Automatic image processing filter generation

2.3.1. Overview

To date, ACTIT and 3D-ACTIT have been applied to various image processing problems and have demonstrated their effectiveness. However, several issues have been identified from these experiments.

- *Image processing filters must be prepared in advance*

In the conventional ACTIT, filters are constructed by combining simple algorithms such as thresholding and neighbourhood processing. In ACTIT, these image processing filters must be prepared in advance before constructing the tree-structured filter. If the features of the image are known, filters utilising those features can be added. However, this requires preparing new filters each time the target changes. Furthermore, it is possible to prepare all conceivable filters, including those with complex algorithms or those that limit the processing targets. However, this expands the search space, making it difficult to construct effective algorithms. Additionally, using filters with complex algorithms or those that limit the processing targets increases the likelihood of constructing algorithms that are highly specialised for the training image set. It is necessary to automatically capture the features of the input image and dynamically generate filters.

- *Algorithms such as shape matching through expansion and contraction processing may be constructed.*

When analysing the algorithm of the constructed tree-structured filter, it was often observed that in the first half of the process, the target object was extracted from the pixel value conversion process, and in the second half, the image was brought closer to the target image using a puzzle-like algorithm such as expansion and contraction processing after binarization. Even without using binarization processing, most pixel value conversion filters can only perform predetermined processing and do not consider pixel value features, resulting in a combination problem aimed solely at making the output resemble the target image. In other words, the current image processing filter set may construct algorithms that do not consider the pixel values or their features of the original image, as long as the final result resembles the target image. Such algorithms are highly dependent on the size and shape of the target object rather than its characteristics such as texture, making it difficult to adapt to changes in shooting conditions or positions in medical images. To improve the generalisability of the algorithm to testing images, it is necessary to consider not only the size of simple pixel values but also its characteristics of the original image and the image being transformed.

To address these issues, it is necessary to develop an algorithm that automatically captures the features of the input image, dynamically generates filters, and performs image conversion

with always considering the pixel value characteristics of the original image and the image data during conversion. In this section, we describe a proposed method for capturing image features and automatically generating filters. In medical image processing, there are individual differences in the size and position of organs, therefore we focus on the texture patterns of each organ and tissue that are not affected by these individual differences, and propose an image processing filter automatic generation method for an image processing filter combination method that takes these into account.

2.3.2. Feature calculation in sliding window

In this method, rather than calculating a single image feature from the entire image, local features are calculated for each pixel to capture the characteristics of each tissue. A window containing a pixel of interest and its surrounding pixels is scanned across the image, and the features within the window are calculated with shifting the pixel of interest one pixel at a time. In this method, texture features were primarily prepared to capture the local regional characteristics of each organ and tissue in medical images. By considering the pixel value characteristics of the surrounding pixels along with the target pixel, this method provides clues for organ identification and tissue classification that were difficult to obtain using simple pixel value magnitudes alone.

Figure 22 shows an example of local features calculated from a PET tomographic image and visualised. The Feature Image is a visualisation of the features calculated using the window. One image is output for each feature. In this image, higher feature values are represented in red, and lower values in blue. As can be seen from Figure 22, this method reveals features that are different from those in the original image.

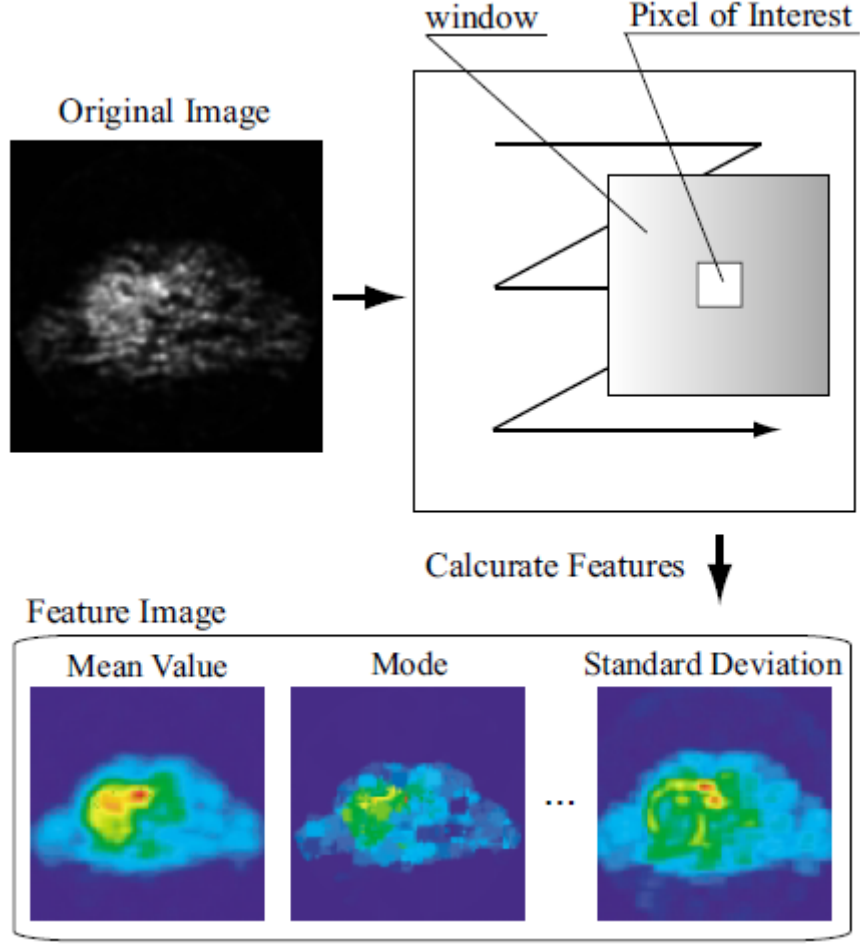


Figure 22 Feature calculation using sliding window

2.3.3. Image transformation using calculated features

Image conversion is performed using the features calculated using windows as in the previous section. The image conversion method determines the conversion rate based on the high or low values of the calculated features. Specifically, as shown in Figure 23, the feature image calculated is overlaid on the original image, and the pixel values of the original image are strengthened or weakened depending on the feature values at the same positions. The conversion formula is shown below.

$$Output = \sum_{i=1}^W \sum_{j=1}^H Input_{ij} \cdot (Feature_{ij} \cdot Weight + Offset) \quad (3)$$

Here, *Output* is the output image, *Input_{ij}* is the input image, and *Feature_{ij}* represents the feature image such as ones shown in Figure 22. *Feature_{ij}* takes from 0.0 to 1.0. To adjust the conversion rate, change the values of *weight* and *offset*. In the experiments in this section, *weight* = 1.0 and *offset* = 0.5 are fixed. This method enables the automatic generation of image

processing filters that capture features of the input image by simply prepared feature quantities calculated within the window, with maintaining the simplicity of the original algorithm. This is achieved by replacing the image processing filters which previously followed simple rules independent from image position and pixel value, with a single filter. Additionally, this method applies to tree-structured filters, where image features are calculated and filters are created for each image, whether it is the original image or an image undergoing transformation. In other words, the filter can take an infinite number of forms depending on the input image.

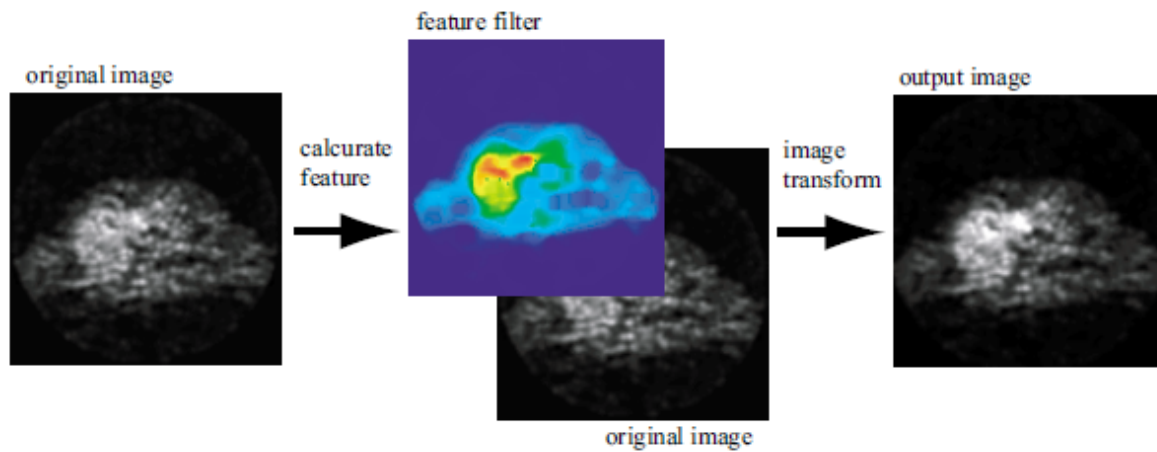


Figure 23 Image processing filter considering features

2.3.4. Integration of automatic image processing filter generation into ACTIT

We implemented an automatic generation method for image processing considering features in ACTIT and automatically constructed image processing. To compare the proposed method with conventional filter sets, we conducted an experiment to extract lung regions from PET tomographic images. Figure 24 shows the training image set used in this experiment. As shown in Figure 24, the target image is filled with white for the lung region and black for other regions. The weight image was created such that the weight of the lung region is higher than that of other regions in the fitness calculation. Using this training image set, experiments were conducted with both the conventional filter set and the proposed filter set, and comparisons were made of the output images and the constructed algorithms.

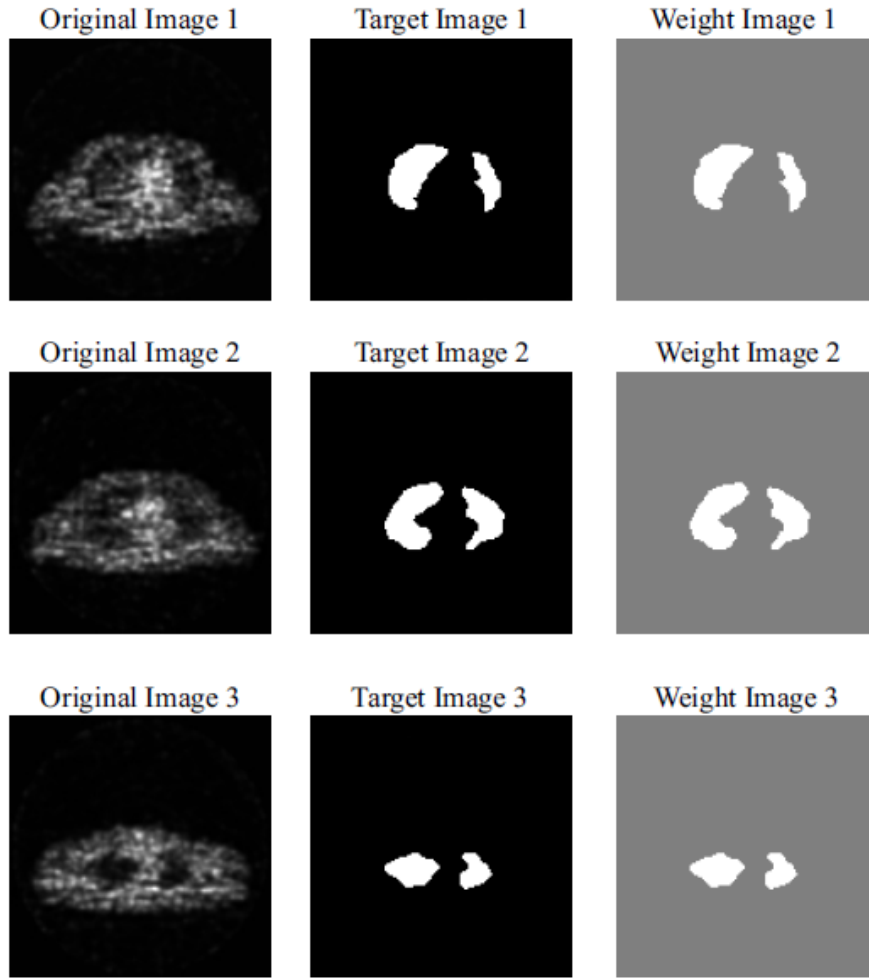


Figure 24 Training data for extracting lung region from PET images

The feature filters prepared for this experiment are shown in Table 6. Two types of filters, "strengthen" and "weaken" were prepared for each feature. The "strengthen" filter normalises the calculated feature values and applies the normalised values as it is in the transformation followed by equation (3). The "weaken" filter normalises the calculated feature values, inverts them, and then performs the transformation. In other words, with the "weaken" filter, lower feature values are emphasised in the grayscale values. The first character of the symbol If the first letter is uppercase, it represents the "enhance (strengthen)" filter. If it is lowercase, it represents the "weaken" filter. Additionally, the window size used for feature calculation was empirically set to two types: 9×9 pixels and 13×13 pixels. The last digit of the symbol indicates the window size: "1" represents 9×9 pixels, and "2" represents 13×13 pixels. In other words, four types of filters were prepared from a single feature. The proposed method, ACTIT, uses the filters in Table 6 for 1-input 1-output filters and the filters in Table 1 for 2-input 1-output filters. Using these filters, the lung region is extracted from PET images and compared with the conventional ACTIT.

Table 6 Image processing filters considering features and their symbols in ACTIT

9×9	13×13	Filter Name
W-1	w-1	Mean value
WM1	wM1	Maximum value
Wm1	wm1	Minimum value
WS1	wS1	Standard deviation
WT1	wT1	Maximum threshold surface – minimum threshold value
WO1	wO1	1σ probability
WO2	wO2	3σ external probability
WF1	wF1	First quartile
WD1	wD1	Median
WQ1	wQ1	Third quartile
WM2	wM2	Maximum value
Wm2	wm2	Minimum value
Wk1	wk1	Skewness
WL1	wL1	Kurtosis

Experiments using conventional ACTIT

First, Figure 25 shows the tree-structured filter constructed in the experiment using the conventional filter set, the intermediate result images during conversion, and the final output images. As shown in Figure 25, we were able to obtain an algorithm that converts and synthesises the original image input from the leaf nodes of the tree and produces output images that are very close to the target images. Looking more closely at the tree-structured filter algorithm, we first take the algebraic product "AIP" of the image obtained by applying the mean filter "Mea" and the maximum filter "Max" with the original image, thereby widening the voxel value difference between the weakly aggregated lung regions and other regions. Next, the "DrS" filter extracts part of the lung region from the human body using the "DrS" filter, and the "SAB" filter converts the black background areas to white, leaving only the lung region blocks. In the latter half of the processing, noise is removed using the "Mea" filter, and the image is expanded using the "Exp" filter, resulting in an output image similar to the target image. However, this algorithm suffers from the same issue discussed in Section 2.3.1, where the latter half of the processing involves shape matching due to expansion processing. Conventional filter sets include filters that apply drastic changes, such as binarization or area thresholding. When such filters are incorporated into the early or middle stages of processing, pixel value information in the image is lost, limiting the filters available for use in the latter half of the processing. Furthermore, previous experiments have shown that overusing thresholding processing reduces the generalisability of the algorithm to testing images which are not used in training, it means tree structural filter construction. However, simply removing such filters from the filter set makes it difficult to construct image processing for the training image set using only the remaining image processing filters, creating a trade-off. While it is possible to manipulate the order of filters during training, manipulating the order or combination of nodes in GP evolution significantly restricts freedom and is expected to

degrade GP performance. To solve this problem, a significant modification of the filters is necessary.

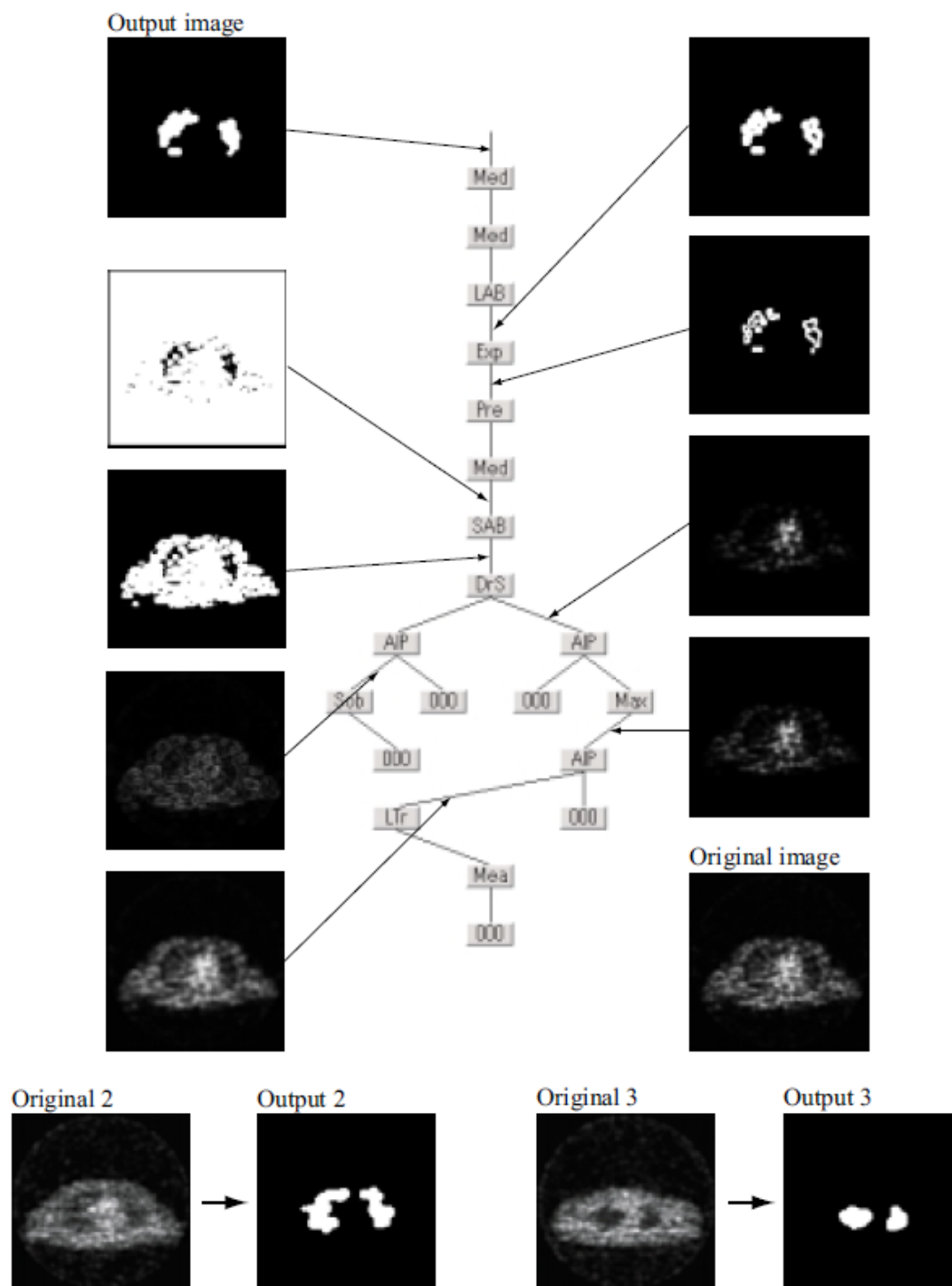


Figure 25 Experimental result using conventional ACTIT

Experiments using ACTIT installing automatic image processing filter generation method

The tree-structured filter constructed using the proposed method in the ACTIT experiment, the intermediate results of the conversion process, and the final output image are shown in Figure 26. Looking at the tree structure in Figure 26 from the bottom, an entire body region mask is created using the intense sum "DrS," and background noise is removed from the difference between the original image and the mask, leaving only the human body as pixel value data. From the middle of the tree structure onwards, image processing filters generated using the proposed method's features are employed. As can be seen from Figure 26, image processing filters that emphasise the lung area are generated by calculating values such as the first quartile "wF2" and the median "WD2" from the incoming image data. Gradual pixel value conversion is performed using the image processing filter generated from the features, resulting in an output image where only the lung region has high pixel values. The output image after binarization shows results very close to the target image, similar to the experiments using the conventional ACTIT method. The algorithm constructed using the proposed method performs image conversion while considering texture features calculated from tonal images, which is clearly different from the binary image matching performed by the conventional filter set. The expansion processing of binary images in the latter half of the processing in Figure 25 causes the loss of pixel value information in the image, and therefore, this processing is intended to increase the fitness rather than considering the characteristics of the image. On the other hand, as shown in Figure 26, the proposed method consistently performs transformations that take into account the characteristics of the image. Although it is difficult to construct an algorithm that brings about a rapid increase in fitness due to the absence of filters that perform dramatic transformations, as in conventional filter sets, the proposed method steadily and reliably transforms images to bring them closer to the target image, enabling the construction of a convincing algorithm.

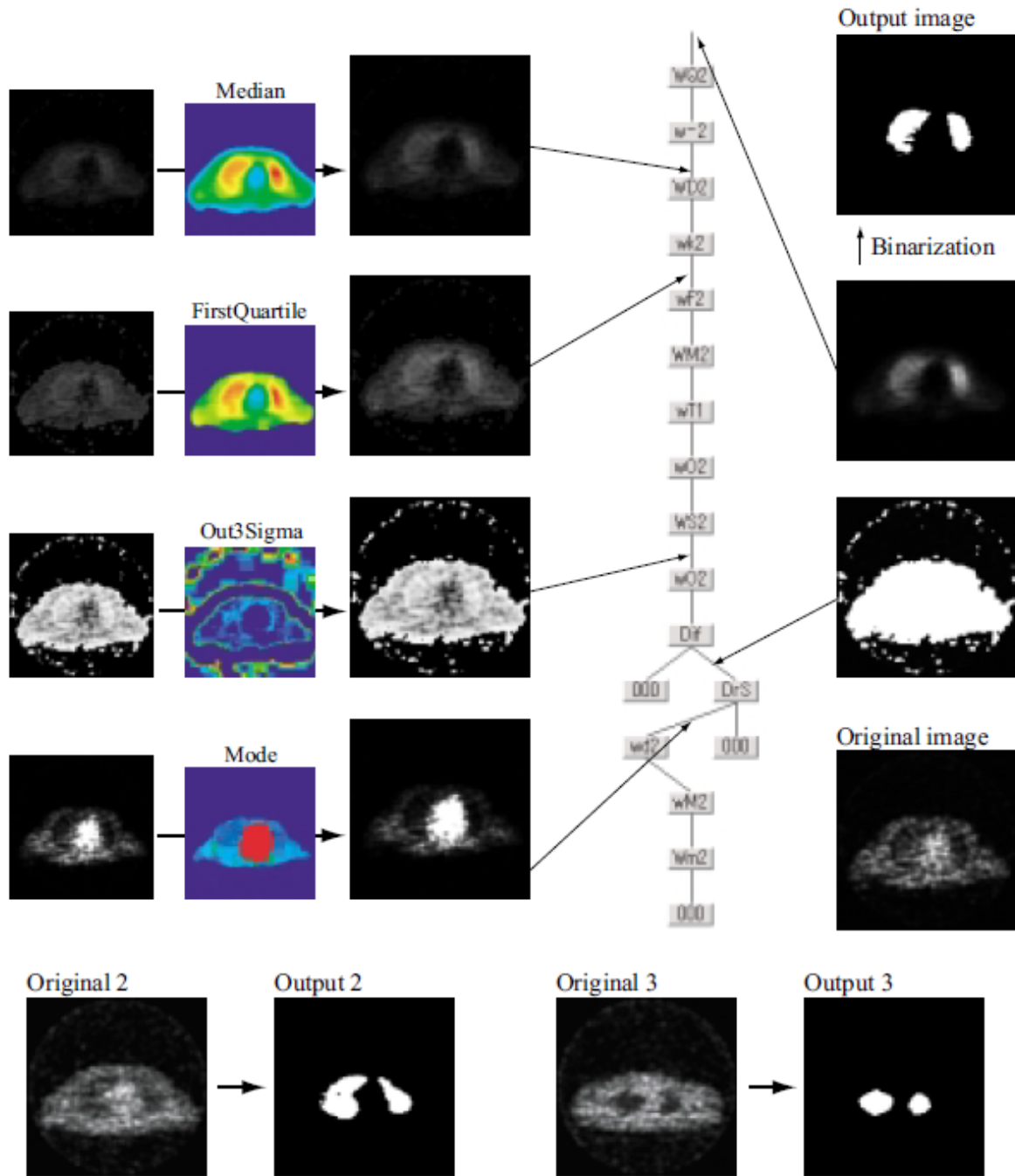


Figure 26 Experimental result of ACTIT using proposed filter generation method

Application to testing data

The results of applying the constructed tree-structured filter to testing images are shown in Figure 27. To be more specific, Figure 27 shows the correct examples obtained by manually marking the lung regions in testing images, the experimental results using the conventional filter set, and the output images obtained by applying the tree-structured filter constructed using the image processing filter automatic generation method. The adaptability of each output image is calculated based on the manually created correct examples. As can be seen

from the results, the proposed method demonstrates higher adaptability across all images. Additionally, visually, while the output images from the conventional filter set exhibit over-extraction and under-extraction, the proposed method achieves extraction that captures the shape of the lungs with minimal noise.

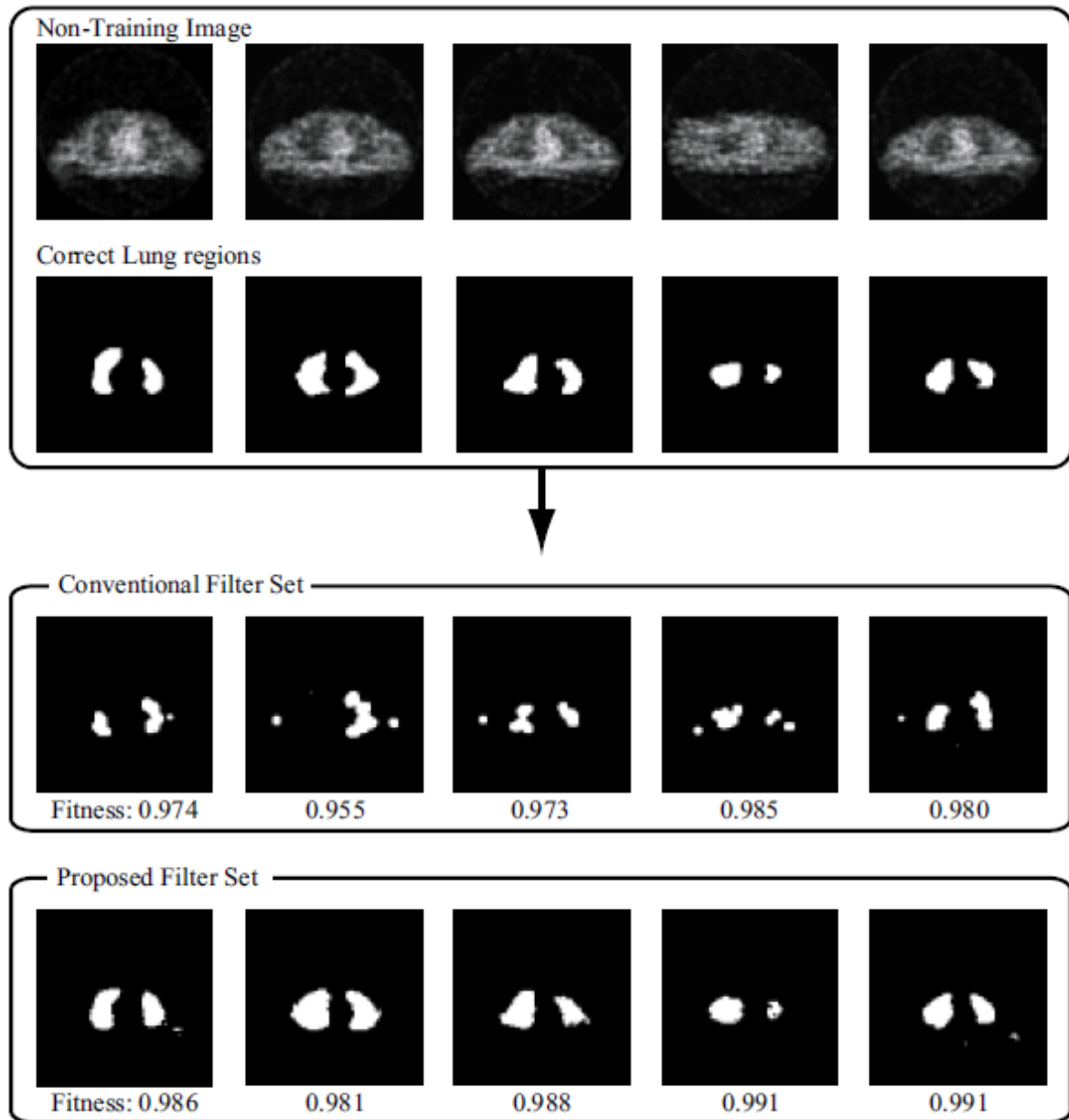


Figure 27 Results on testing images

Experiments using other medical images

To verify the effectiveness of the automatic image processing filter generation method, we applied it to medical images other than PET. The goal of this experiment was to construct an image processing method for segmenting the liver region from abdominal CT images. The training image set used in this experiment is shown in Figure 28. The CT images in Figure 28

contain organs such as the liver, kidneys, pancreas, and spleen. The researches on segmenting each organ from CT images containing multiple organs are actively being conducted. Most of these studies use a probabilistic atlas of the human body to perform segmentation based on the presence of organs. In this experiment, we aim to achieve segmentation processing using only the combination of filters implemented in ACTIT, without using any probabilistic atlas or information such as the size or position of each organ. Figure 29 shows the tree-structured filter constructed from the training image set, the intermediate results of the conversion, and the final output image. Analysing the tree structure from the bottom, we first weaken the grey scale values of the kidneys using features such as the mode and median. Next, we create an image in which the target liver is slightly highlighted by taking the difference between the images converted by two subtrees. In the latter half, filters that further highlight the liver, such as the third quartile and mode, are generated. Although the pancreas, which has similar pixel value characteristics to the liver, remains slightly visible, the extraction of the liver region is successful.

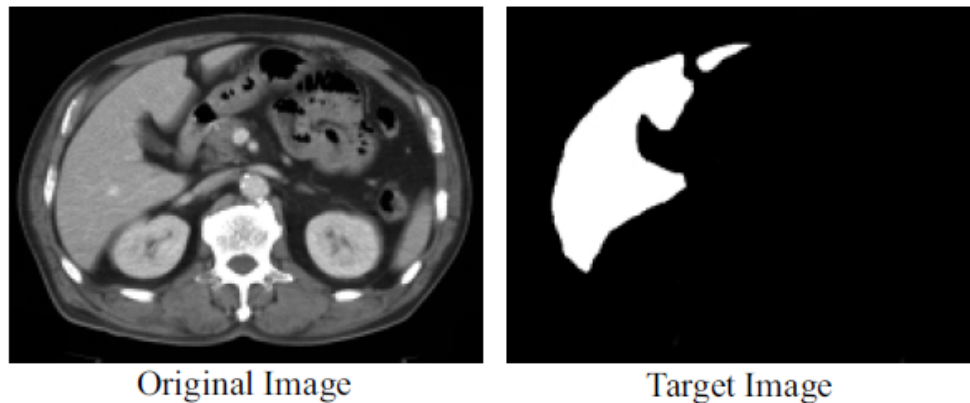


Figure 28 Training data for extracting liver region from CT image

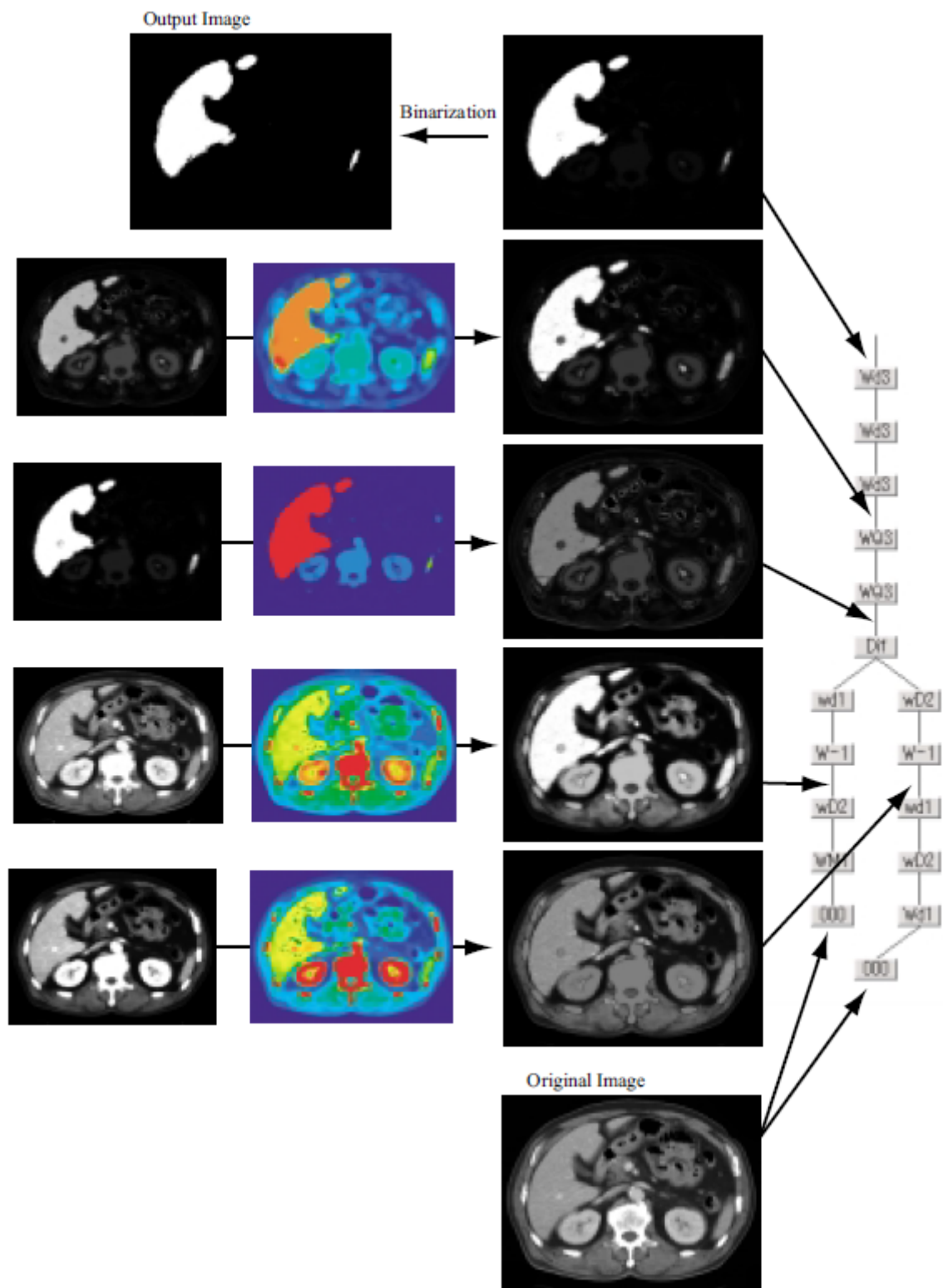


Figure 29 Result on training data for liver segmentation by using automatic filter generation method

2.3.5. Integration of automatic image processing filter generation into FFGIN

Implementation of an image processing filter automatic generation method considering features in FFGIN. FFGIN optimises both the structure and parameters of filters by modifying the genotype. To date, Parameter Tunable FFGIN (PT-FFGIN), which simultaneously optimises parameters, has been proposed, and its effectiveness has been demonstrated. When implementing the proposed method in PT-FFGIN, we modified the genotype as shown in Figure 30 so that the "window size" used in calculating the features and the weight used in equation (3) when multiplying the calculated features can be optimised simultaneously. Parameter1 is assigned to the window size, and parameter2 is assigned to the weight. The window size is determined by the value of parameter1 as follows: 7×7 , 9×9 , 11×11 , 13×13 , and 15×15 pixels. The weight is adjusted within the range from 0.0 to 1.5 based on the value of parameter2. In this method, the offset in equation (3) is set to 0.3, therefore the pixel values of the input image are multiplied by from 0.3 to 1.8 times in a single filtering operation.

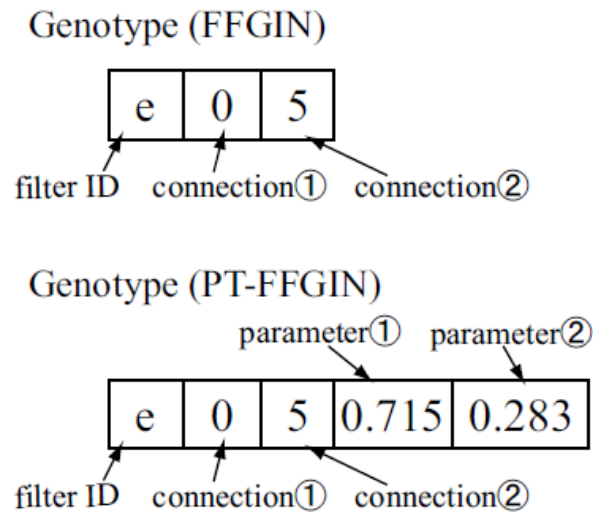


Figure 30 Genotype and phenotype in PT-FFGIN

The prepared filters are shown in Table 7. Since the window size is optimised, two types of filters, "strengthen" and "weaken," are prepared for each feature. As in Section 2.3.4, if the first letter of the symbol is uppercase, it represents a "strengthen" filter, and if it is lowercase, it represents a "weaken" filter. Additionally, to effectively utilise partial structures when constructing multi-output algorithms, gamma correction filters and inversion filters have been added. We believe that the addition of these filters will enable us to further highlight features and construct partial structures that generate images that act as masks through inversion processing. The gamma correction filter determines the gamma value in the range from 0.2 to 2.2 based on parameter1 in the genotype in Figure 30. No parameters are used for the inversion filter. Image processing is constructed using these filters.

Table 7 Image processing filters integrated into PT-FFGIN

1-Input / 1-Output Filters	
W- / w-	Mean value
WM / wM	Maximum value
Wm / wm	Minimum value
WS / wS	Standard deviation
WT / wT	Maximum threshold surface – minimum threshold value
WO / wO	1σ probability
WO2 / wO2	3σ external probability
WF / wF	First quartile
WD / wD	Median
WQ / wQ	Third quartile
Wd / wd	Mode
Wk / wk	Skewness
WL / wL	Kurtosis
G2	Gamma correction
I2	Inversion
2-Input / 1-Output Filters (Input image 1: f_1; Input image 2: f_2)	
2L	Logical AND: $\min(f_1, f_2)$
2I	Logical OR: $\max(f_1, f_2)$
2A	Arithmetic product: $f_1 \times f_2 / V_{max}$
2a	Arithmetic sum with normalization: $f_1 + f_2 - f_1 \times f_2 / V_{max}$
2B	Bounded sum: $f_1 + f_2 - V_{max}$
2b	Bounded product: $\max(f_1 + f_2 - V_{max}, 0)$
2u	Inhibition: $f_1 = 0 \rightarrow f_2 = 0 \rightarrow f_1$, else 0; $f_2 = 1 \rightarrow V_{max}$
2U	Inhibition (inverse): $f_1 = V_{max} \rightarrow f_2 = V_{max} \rightarrow f_1$, else 0
2D	Difference filter: $ f_1 - f_2 $
21	Pass-through: Output input image 1 as is
22	Pass-through: Output input image 2 as is

Simultaneous segmentation of multiple regions from a PET image

ACTIT combines image processing filters in a tree structure, resulting in a single output. However, FFGIN, which combines image processing filters in a feedforward network structure, can have multiple outputs in a single structure. In this section, we conduct an experimental study to construct a process for segmenting multiple regions from PET tomographic images using a single network structure algorithm. By expressing multiple region segmentation processes within a single structure, we can suppress the increase in computational cost that occurs when constructing a tree-structured algorithm for each target region, as in the previous ACTIT. Additionally, by reusing partial structures within a single network structure, the algorithm can be made more compact and can leverage internal interactions. In this experiment, two sets of training images were used. The training image sets used are shown in Figure 31 and Figure 32. Figure 31 was created with the goal of extracting a relatively weakly aggregated lung region, the heart sandwiched between the lungs, and the entire body. The three target images correspond to the lungs, heart, and entire body, respectively. Figure 32

shows training images created using lung cancer patient data. As can be seen from the original images, strong accumulation exists within the lung regions that normally exhibit weak accumulation. This strong accumulation originates from cancer cells. This training image set aims to simultaneously extract normal lung regions and regions showing abnormal accumulation. The target images correspond to the lung, abnormal accumulation, and the entire body, respectively.

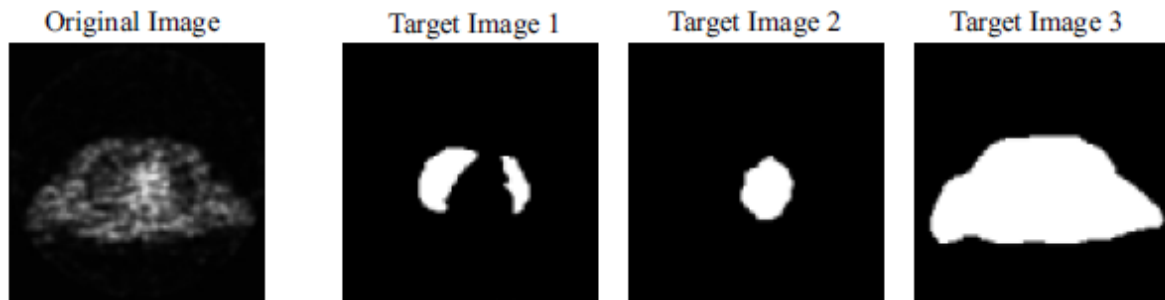


Figure 31 Training data for extracting lung, heart and entire body

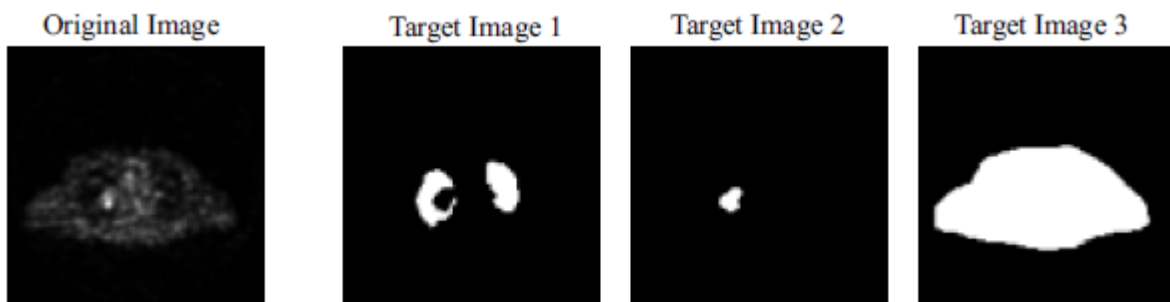


Figure 32 Training data for extracting lung, cancer region and entire body

Multiple region segmentation for lung, heart and entire body

The training image set of Figure 31 was input into PT-FFGIN using the proposed method, and the constructed network structure filter, intermediate results, and final output images are shown in Figure 34. The constructed network structure filter and intermediate results are shown in Figure 34. At first, it should be explained that, as opposed to the tree-structural filter from ACTIT, the processing flow in Figure 34 is from top to bottom, it means input images are fed from top of network structure. Analysing the network structure from the input in Figure 34, we can see that a common structure is used for multiple outputs in the early stages of processing. In the early stages of processing, smoothing is performed using standard deviation "wS" and mean "W-" filters. After applying the inversion filter "I2", the structure branches. One branch applies a maximum filter "WM" to the inverted image to emphasise the background and lung areas, then calculates the maximum value again using a large window of 13×13 pixels, and finally applies a filter "wM" to weaken areas with high feature values, thereby weakening only the background and extracting only the lung area to obtain the output image 1. In the other route, the inverted image is first smoothed using the average filter "w-"

centred on the heart accumulation area, and then processed to unify the entire body with the same intensity value. The result is combined with the output node that extracts the entire body, yielding output image 3. Furthermore, by applying the maximum value filter "WM" to the smoothed image, we captured and emphasised the strong heart accumulation features compared to the surrounding area, resulting in output image 2, which extracts only the heart region. All three output images are close to the target images, and the network structure shows that a very compact algorithm is expressed with reusing a common structure.

Furthermore, the results of applying the obtained network structure-based filter to a testing image are shown in Figure 34. From Figure 34 shows that the constructed algorithm is effective even for testing images.

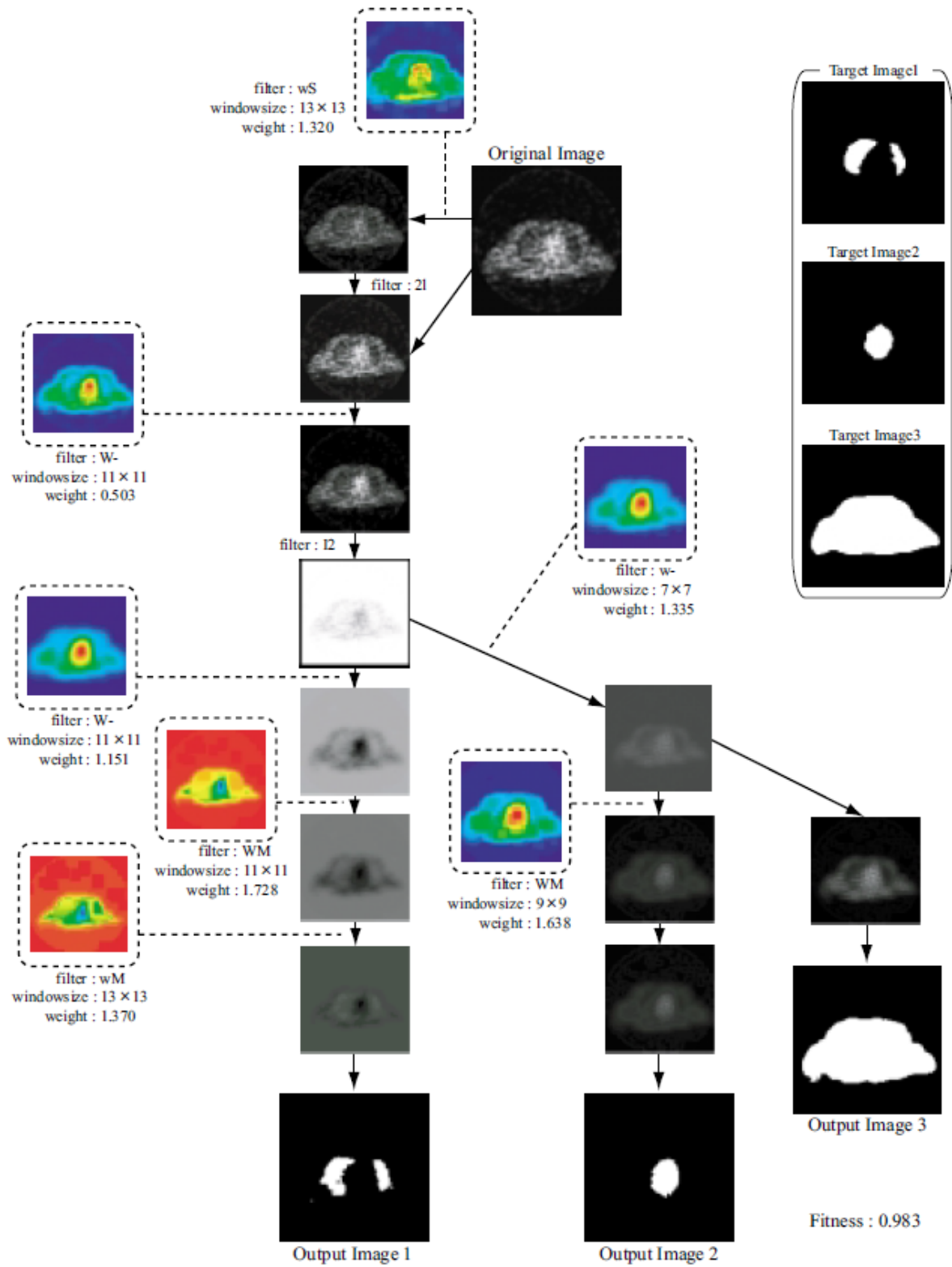


Figure 33 Analysis of obtained network structured filter for segmentation of lung, heart and entire body

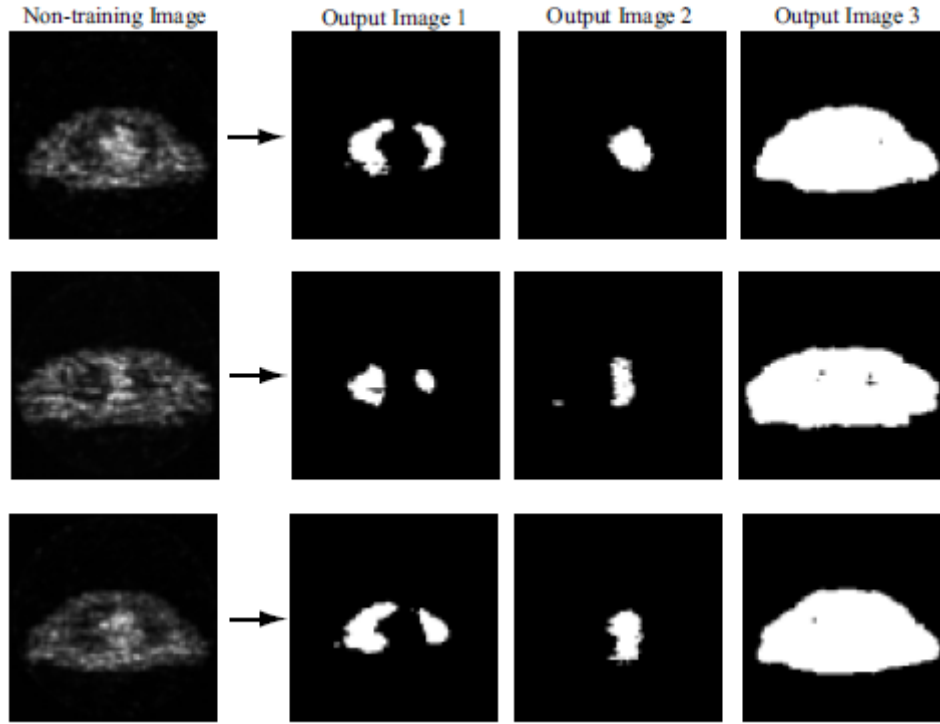


Figure 34 Results on testing data

Multiple region segmentation for lung, cancer region and entire body

The training image set in Figure 32 was input into PT-FFGIN using the proposed method, and the network structure filter constructed, the intermediate results, and the final output image are shown in Figure 35. Same as Figure 33, the processing flow of Figure 35 is from top to bottom. When analysing the network structure filter from the input, it can be seen that the initial filters are common to both structures. First, the difference between the maximum and minimum grayscale values is calculated, and a filter "wT" that weakens areas with high feature values is applied to reduce abnormal accumulations with high grayscale values. Then, by calculating the intensity product "2U" between the image and the original image, only the abnormal accumulations are extracted, as shown in the output image 2. Next, an inversion filter is applied to the image with weakened abnormal accumulations, and filters using the first quartile "wF" and the mean value "w-" are used to weaken the background and lung areas, thereby distinguishing the background from the human body and filling in the weak accumulations within the body (lung areas), resulting in the extraction of the entire body region as shown in output image 3. Finally, we apply an average value filter "W-" to the inverted image to emphasise the lung regions and a maximum value filter "wM" to weaken the background, ultimately extracting the lung regions as shown in output image 1.

The results of applying the obtained network-structured filters to testing images are shown in Figure 36. From the output image in Figure 36, we were able to extract lung cancer accumulations that appear unpredictably within the lung region, with also successfully

extracting normal lung regions and the entire body region. These results demonstrate the effectiveness of the proposed method.

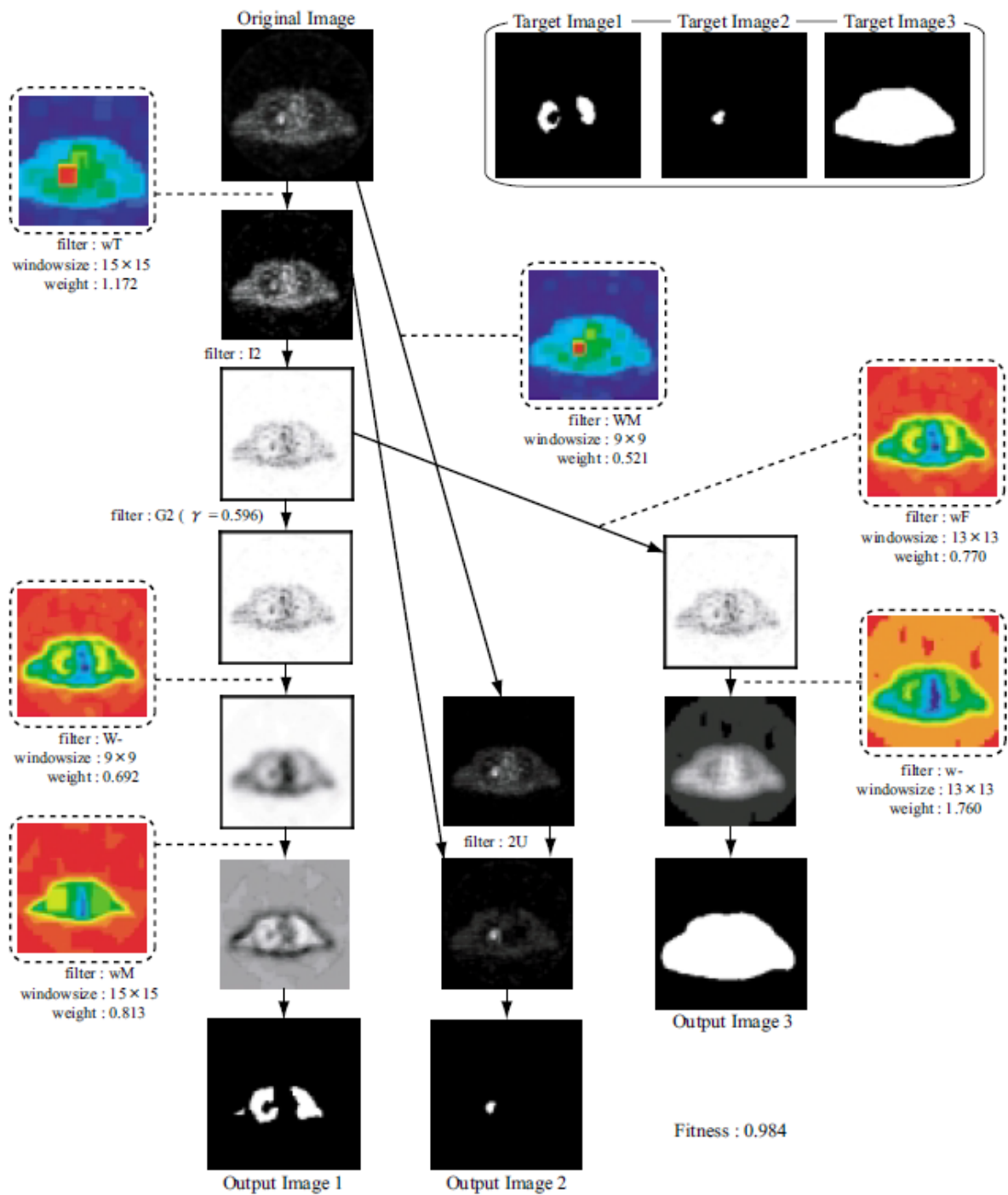


Figure 35 Analysis of obtained network structured filter for segmentation of lung, cancer region and entire body

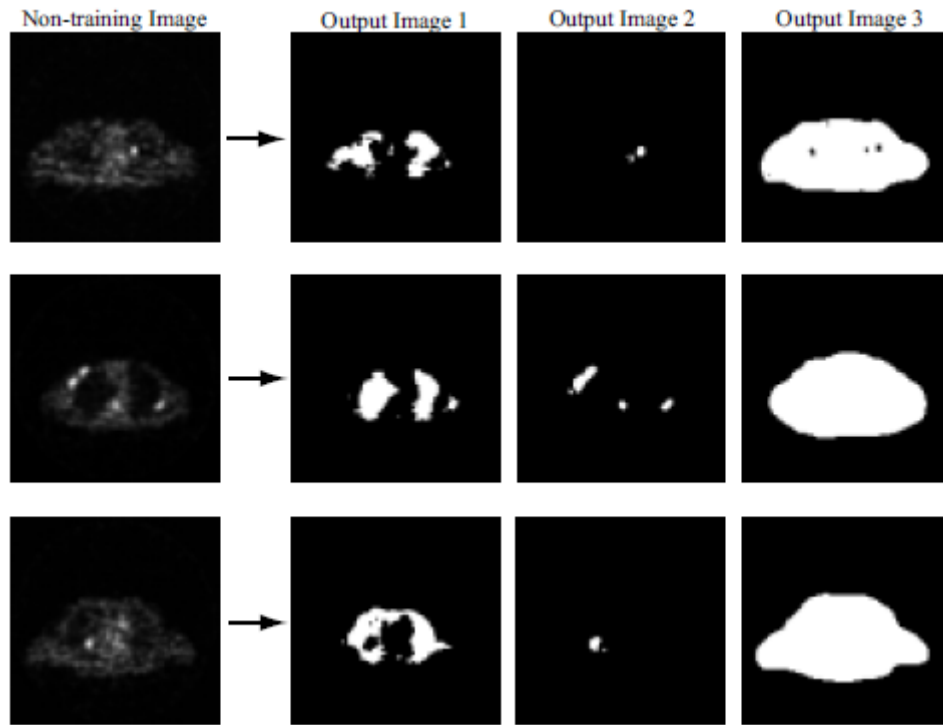


Figure 36 Results on testing data

2.3.6. Summary

In image processing filter combination methods such as ACTIT and GIN, several issues have been clarified, such as the need to prepare a filter set in advance before constructing algorithms, and shape matching with the target image. However, by using an image processing filter automatic generation method that considers local features in images, it is possible to automatically generate image processing filters that capture pixel value features such as texture patterns or statistical features from input images, thereby resolving these issues. Furthermore, this method was found to be suitable for medical image processing, as it enabled accurate segmentation of organs and tissues in PET and CT images, which cannot be simply determined based on pixel value magnitudes, by applying this technique. In the future, we aim to add new features and achieve the construction of more advanced image processing methods.

2.4. Tissue classification using local features

2.4.1. Overview

To date, image processing has been developed using image conversion approaches such as ACTIT and GIN for organ segmentation and abnormal region extraction. However, for functional evaluation of organs from PET images and diagnosis that considers the patient's health status and characteristics, it is necessary to analyse organs and tissues not only based on signal intensity but also by comparing them with normal patterns and analysing texture patterns of grey-scale values. For example, in Alzheimer's disease diagnosis using PET images, it is necessary to consider both the normal signal patterns of the brain and the signal characteristics of the patient to identify regions with impaired function. In such cases, abnormal regions may exist as potential abnormalities, meaning that abnormal patterns are infinite, and it is difficult to explicitly identify the characteristics of abnormal regions or parts. When addressing such issues with ACTIT or GIN, creating target images becomes extremely challenging. Therefore, we propose an image analysis method that captures the characteristics of each tissue. First, as in the previous section, we calculate multiple types of local features using windows. These features are represented as multidimensional vectors for each pixel, and clustering is performed on a pixel-by-pixel basis. For clustering, we apply self-organising maps (SOM). By inputting the multidimensional vectors of each pixel into the self-organising map and training it, pixels with similar features are mapped to nearby locations on the map. This allows us to perform tissue classification based on the map. In this section, we apply this method to brain PET images and abdominal CT images to perform image analysis for Alzheimer's disease diagnosis and tissue classification of various organs.

2.4.2. Image diagnosis for Alzheimer's disease

In today's rapidly aging society, the proportion of elderly people with dementia is increasing year by year, and interest in treatment and diagnostic methods is growing. The most representative types of dementia are vascular dementia and Alzheimer's dementia. Vascular dementia is caused by blockages in the blood vessels of the brain, such as infarcts. On the other hand, Alzheimer's dementia has an unknown cause and does not have identifiable lesions like those seen in vascular dementia. Therefore, when performing imaging diagnostics, it is diagnosed based on findings such as cerebral atrophy or enlargement of the ventricles. Traditional imaging diagnostic techniques have faced numerous challenges in diagnosing Alzheimer's dementia, however, advancements in imaging technology have enabled more detailed differentiation of brain tissue, leading to extensive research on imaging diagnostics for Alzheimer's dementia [33][34][35]. Kodama et al. proposed an objective diagnostic method by calculating ventricular area and hippocampal area from head MRI and comparing them with those of healthy individuals. Additionally, Goto et al. investigated numerical errors arising from differences in imaging methods using the VSRAD (Voxel-Based Specific Regional

Analysis System for Alzheimer's Disease) software, which numerically analyses the degree of hippocampal atrophy from 3D-T1-weighted images. Research on imaging diagnosis using PET and SPECT is also actively being conducted [36]. By using PET or SPECT, it is possible to investigate declines in brain function that are not visible in shape, thereby enabling the detection of early-stage Alzheimer's dementia that is not detectable by CT or MRI. However, functional images such as PET show significant individual variability. Even in the same patient, differences can arise depending on their health condition on a given day. Currently, the diagnosis of Alzheimer's dementia using PET is performed by visual inspection by a physician. Doctors consider a database of healthy individuals in their minds and the patient's health condition on the day of the examination to identify overall and localised declines in brain function. To support such diagnosis, research is being conducted to create a PET database of healthy individuals and identify abnormal areas by matching them. However, this diagnostic method compares results with an average database of healthy individuals, making it difficult to account for individual differences in shape and function, and thus unable to diagnose based on the functional characteristics of the patient in question. Figure 37 shows examples of brain PET images of healthy individuals and Alzheimer's disease patients. As seen in Figure 37, the boundaries are unclear, and even in homogeneous tissues, PET values lack uniformity, making image processing not as straightforward as one might think. There is a need for PET analysis methods that can reduce the burden on physicians with enabling objective and quantitative evaluations.

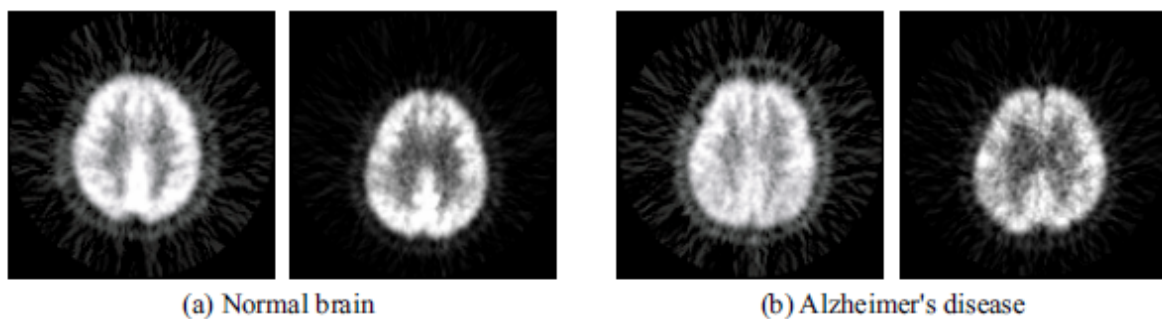


Figure 37 Examples of brain PET images of healthy individuals and Alzheimer's disease patients

2.4.3. Tissue classification using self-organizing maps

Image analysis is performed using texture features that quantify the regularity of intensity variations in images. In this study, we focus on tissue classification and perform image analysis by clustering texture features extracted from brain PET images and abdominal CT images using self-organising maps.

Calculation of texture features

Texture features are quantitative measures of regularity in intensity variations. These texture features have been used in research to perform image segmentation, where simple edge

extraction processes have difficulty determining boundaries, and image classification based on texture characteristics. In this method, as in the previous section, we do not calculate a single texture feature for the entire image, however rather calculate texture features for each pixel. A window including a target pixel and its surrounding pixels is scanned across the image, and the feature values within the window are calculated at each position. Figure 38 shows an example of calculating texture features from a PET cross-sectional image and visualising them. The Feature Image is an image where the feature values calculated using the window are represented with higher values appearing redder. As shown in Figure 38, a mask image can be used if necessary.

Over 30 types of features, such as standard deviation and mean value, are prepared for calculation. From these features, effective ones are selected and input into a self-organising map that models the neural functions of the cerebral cortex. A self-organising map updates the map according to the input signals, creating a mapping space where signals with similar features gather at the same or nearby nodes on the map. A detailed explanation of self-organising maps is provided in the next subsection. We verify whether tissue classification can be performed using multidimensional texture features.

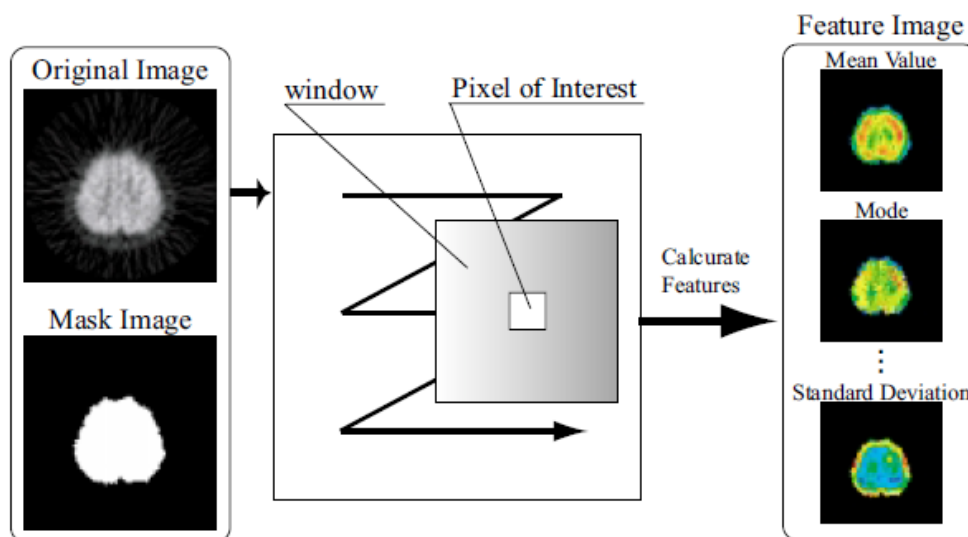


Figure 38 Calculation of texture features using sliding window

Self-organizing map

Self-Organising Maps (SOM) are a type of neural network that models the neural functions of the cerebral cortex. It is a type of unsupervised learning method, and it updates the map in response to input signals and creates a mapping space where signals with similar characteristics are grouped together. First, an initial map is created randomly. Next, input vectors are placed at the locations on the map that are closest to their features, and the vectors in the vicinity of the input vectors are updated to move closer to the input vectors. This process is repeated for all input vectors a specified number of times to create the map.

As a result, a mapping space is created where vectors with similar features are grouped together. Figure 39 illustrates the algorithm for a 2D-SOM. By updating the map as shown in Figure 39, multidimensional vectors can be mapped onto a 2D map. Other variations, such as 1D-SOM, torus-shaped SOM, and spherical SOM, have been studied, and their effectiveness has been demonstrated.

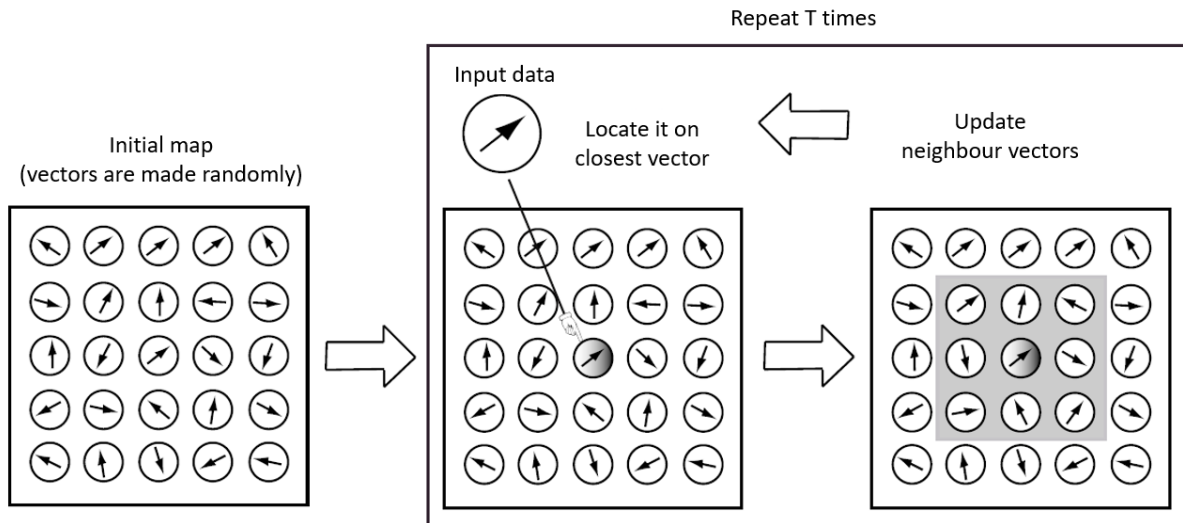


Figure 39 Algorithm of self-organizing map

Tissue classification using self-organizing map

We perform tissue classification of PET images using this method. The specific algorithm is shown in Figure 40. First, multiple texture features are calculated from a window containing the target pixel and its surroundings. This texture feature calculation is performed for each pixel in the image. When the pixel of interest is at the edge of the image and pixels in the window are missing, the features are calculated using only the existing pixels. Similarly, when using a mask image, only the pixels within the mask are used to calculate the features. Next, a self-organising map is created using the calculated multidimensional features as input. After the map has been updated the number of times specified by the user, as described in Figure 40, each pixel is coloured according to its coordinate in the map, and the image is restored. In other words, pixels mapped to the same node in the map are represented by the same colour, and since the colours in the map are graded, pixels closer to each other are represented by colours closer to each other, resulting in an output image. Using this method, it is possible to confirm the similarity of features by colour and analyse images. Using this method, Section 2.4.4, we analyse PET images of healthy individuals and Alzheimer's disease patients, and in Section 2.4.5, we perform tissue classification of abdominal CT images.



Figure 40 Tissue classification using self-organizing map

2.4.4. Experiment of tissue classification on brain PET image

Texture features were extracted from brain PET images with a resolution of $128 \times 128 \times 128$, and tissue classification was performed using the self-organising map. The features used in this experiment were 10 types: mean, standard deviation, difference between maximum and minimum values, first quartile, median, third quartile, mode, skewness, kurtosis, and discriminant analysis threshold. The features were selected manually with the advice of a physician. The window size was empirically set to 7×7 . Figure 41 shows the 10 types of texture features. Figure 41(a) shows the image data of healthy subjects, and Figure 41(b) shows the image data of Alzheimer's disease patients. The mask image used to calculate the features was binarized using the threshold determined by discriminant analysis and labelled to remove small areas. In this experiment, features were calculated from five PET cross-sectional images of healthy individuals and five PET cross-sectional images of Alzheimer's disease patients, and then input simultaneously to create a single map.

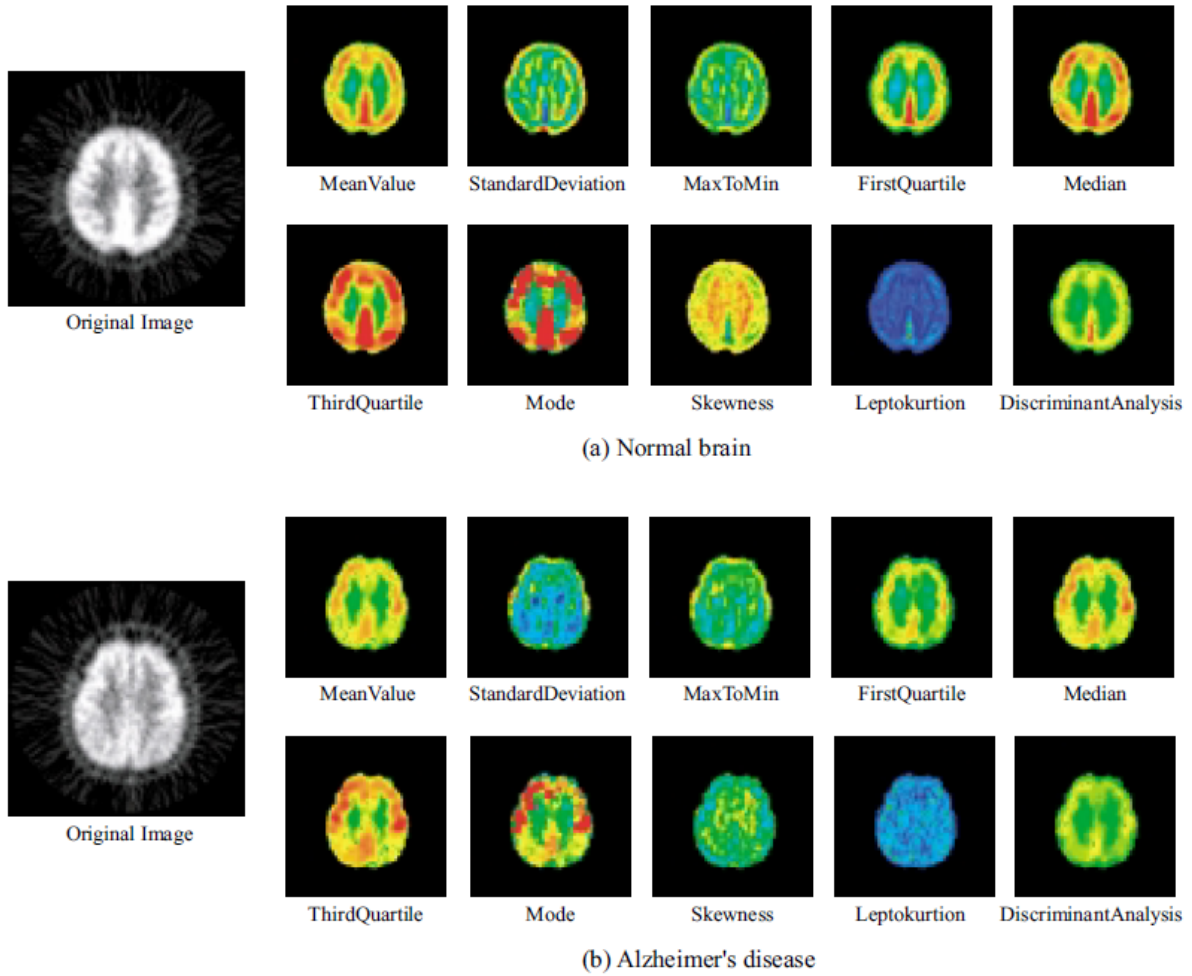


Figure 41 Features used in experiments

The features in Figure 41 were mapped onto self-organising maps using 16x16 resolution. The output images obtained in this experiment are shown in Figure 42 shows the results. The colour of each pixel in the output image corresponds to the colour on the map, and pixels mapped to the same location on the map are represented by the same colour, while pixels mapped to nearby locations on the map are represented by colours that are closer to each other. Looking at the output images, the images of healthy individuals showed connections and symmetry in colour, while the images of individuals with Alzheimer's disease showed scattered colours, with asymmetry observed particularly in the occipital lobe. The classification results obtained by analysing the map in detail are shown in right hand side of Figure 42. When the pixels mapped to each of the 16x16 nodes are represented, differences between healthy individuals and Alzheimer's disease patients appear in the upper part of the map. In the healthy subject data, from the left side of the upper row, there are nodes with little to no mapping, followed by the primary sensorimotor cortex, the parietal association cortex, and the posterior cingulate gyrus and precuneus. In the row below, the left side shows mapping of the ventricles and white matter, indicating good tissue classification. In contrast, in the Alzheimer's disease patient data, the upper row on the left shows areas with reduced signal intensity, and to the right, the primary sensorimotor cortex and the parietal association

cortex are mapped, indicating a mixture of various tissues. Additionally, the overall symmetry of the classified tissues is not observed.

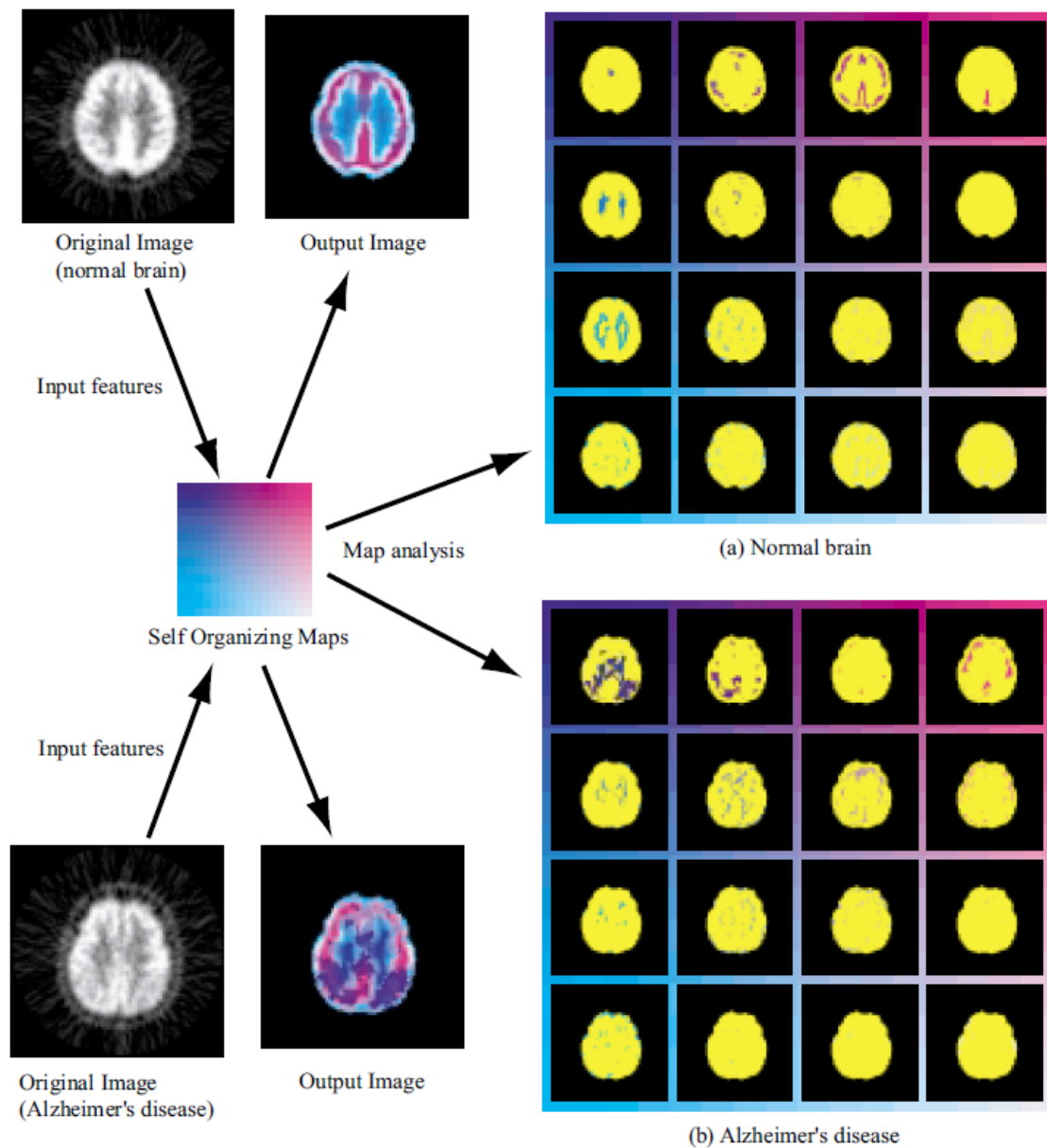


Figure 42 Brain PET tissue classification results

Texture features were extracted from images not used for training, and the results of mapping these features onto the created map are shown in Figure 43. In this figure, Figure 43(a) shows the results mapped from healthy subject data, and Figure 43(b) shows the results mapped from Alzheimer's disease patient data. In the healthy subject data, as in Figure 42, primary sensory motor areas, parietal association areas, posterior cingulate gyrus, precuneus, and ventricles are classified on the map. In Alzheimer's disease patients, regions with reduced signal intensity in the upper left of the map were mapped, and various tissues were mixed in other nodes, showing the same trend as in Figure 42. We believe that this map can be used for brain tissue classification and for detecting tissues suspected of Alzheimer's disease.

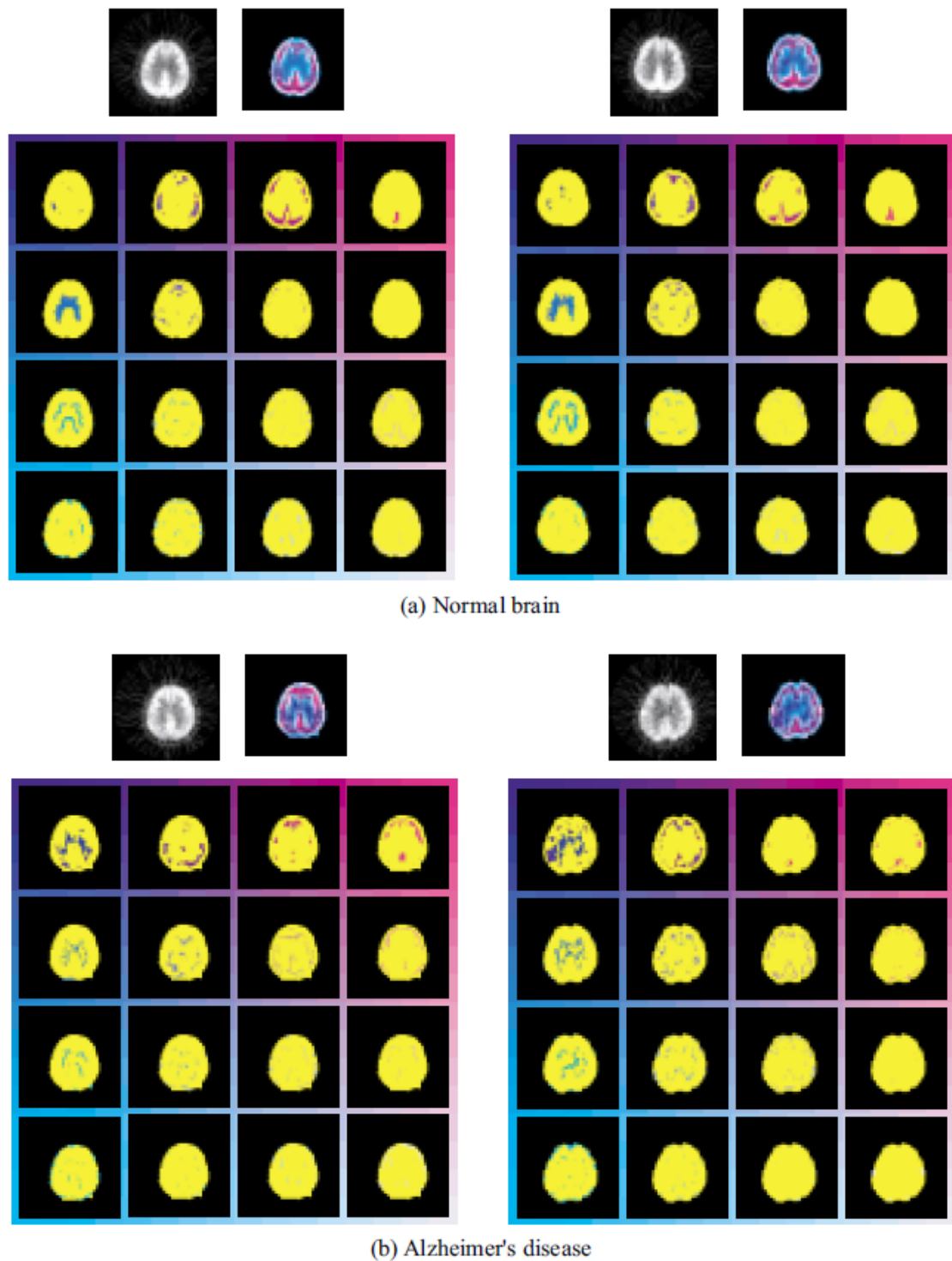


Figure 43 Application to testing images

2.4.5. Tissue classification experiment using CT image

Texture features are extracted from abdominal CT images and tissue classification is performed using self-organising maps. While research on organ segmentation from CT images using methods such as region growth and probabilistic atlases is actively underway, this

experiment aims to classify organ tissues based on texture features extracted from images. The features used in this experiment are five types: mean, first quartile, median, third quartile, and mode. Based on the correct examples of organ regions created with the advice of a physician, we manually selected features that are considered effective for extracting the kidney, liver, and pancreas. Additionally, the window size for feature calculation was empirically set to 15×15 . The original image, the five types of texture features converted into images, and the output results are shown in Figure 44. Although some noise are present, each organ is mapped to distinct nodes on a single map, indicating that tissue classification based on texture features is feasible.

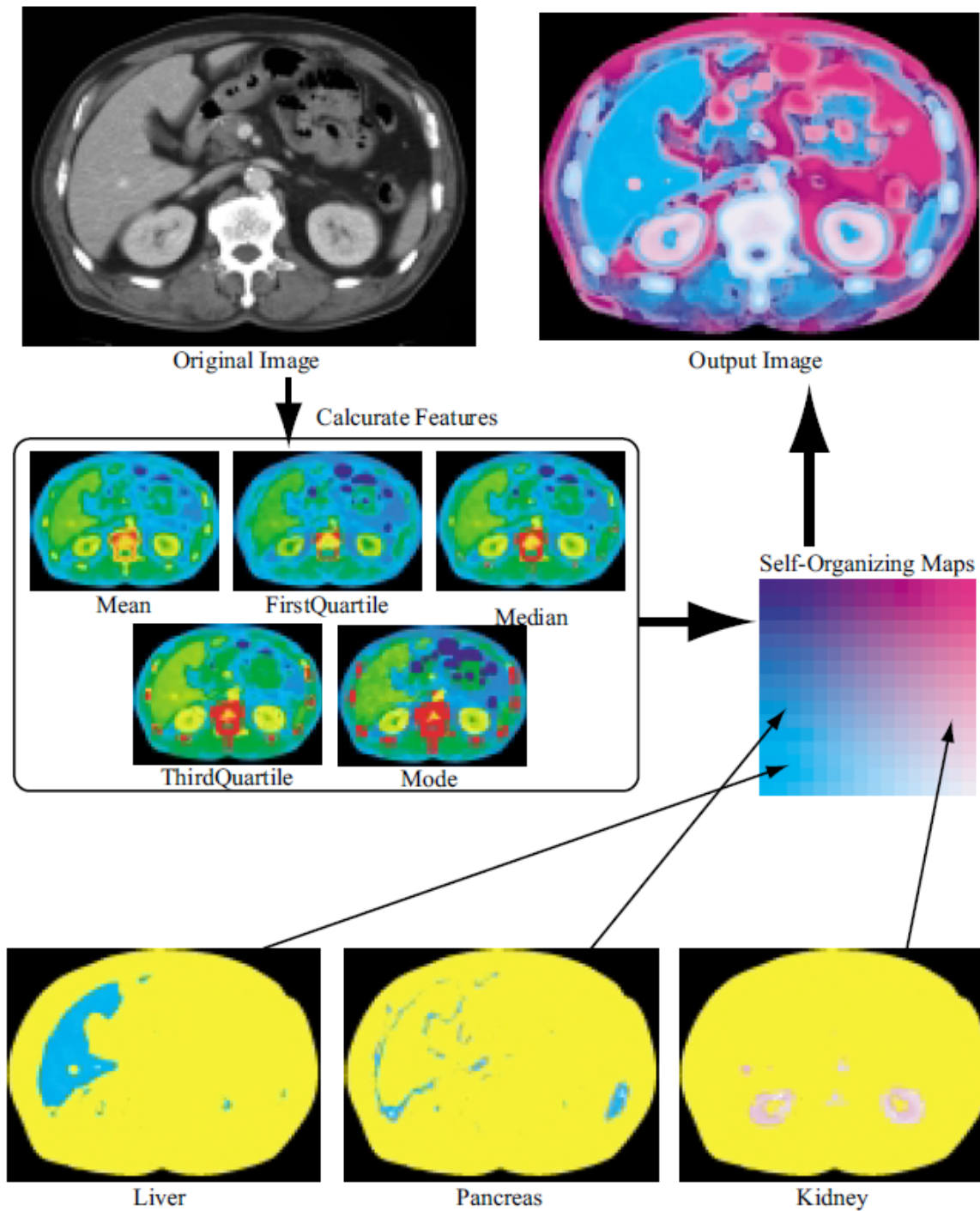


Figure 44 Tissue classification of CT image

2.4.6. Summary

By mapping texture features calculated using windows onto a self-organising map, we were able to perform tissue classification of brain PET images. In particular, we successfully classified primary sensory motor areas and parietal association areas, which serve as landmarks for physicians diagnosing Alzheimer's disease. Even in PET images that were

difficult to analyse using simple grayscale processing, we were able to obtain useful information for analysis by considering features within local regions using windows. In addition, in experiments with CT images, we were able to classify tissues using only texture features without using any knowledge databases such as training data or probability atlases, demonstrating the effectiveness of this method. Currently, the window size and feature selection for this method are determined empirically. In the future, we aim to formalise the evaluation of the output and improve the method to automate the selection of parameters and features. Furthermore, since this method is not limited to PET and CT images but can be applied to all types of images, we plan to analyse medical images other than those used in this experiment.

Chapter 3

Precise Contour Detection and Anormal Pattern

Recognition in Medical domain

2.1. Introduction

First of all, I have to explain that, from this chapter onward, the description of research activities is subject to restrictions due to confidentiality considerations. Therefore, this chapter is explained by using only the information in this chapter have already been disclosed in public such as patents or publications. Consequently, certain sections may appear concise or somewhat abstract, and this limitation should be acknowledged in advance.

Following the completion of the PhD degree, the author joined the research division of Canon Inc. and continued research and development in the field of medical image processing. During this period, the following research topics were addressed:

- Retinal layer segmentation in Optical Coherence Tomography (OCT)
- Organ segmentation (e.g., liver, spleen) in Computed Tomography (CT)
- Lymph node segmentation in CT
- Development of a nodule management system

These studies aimed to advance diagnostic support technologies based on medical image information, with a particular emphasis on the application of mathematical models and artificial intelligence (AI). Specifically, the objective was to delineate organ boundaries and pathological regions in tomographic images such as CT with high precision. In contrast to the methods introduced in Chapter 2, the research activities described here focused on precise extraction. As discussed in later subsections, the primary objective was the quantitative assessment of organs and lesions, which required pixel-level accuracy in classification. To address this requirement, a two-stage strategy was adopted: AI-based coarse extraction at the initial stage, followed by the application of mathematical models, such as the Level Set method [37] and Graph Cut method [38], for fine segmentation.

In 2016, the author transitioned from Canon Inc. to MEDIAN Technologies in France, where the research focus remained on the detection and segmentation of lesions in medical images for radiological practice. The principal topics investigated during this period included:

- Non-Alcoholic Steatohepatitis (NASH) assessment based on texture analysis of liver regions in CT images
- Benign/malignant classification of lung nodules using deep learning

MEDIAN Technologies has long collected data from numerous hospitals and established a large-scale medical imaging database. This enabled extensive experimentation with a variety

of pattern recognition approaches, including deep learning [39], clustering methods, and other machine learning techniques, in the context of diagnostic support.

In the subsequent subsections of this chapter, the objectives of these research projects and the technical approaches employed will be described in detail, focusing primarily on information available from patents and other publicly accessible sources.

3.2. Retinal layer segmentation in Optical Coherence Tomography (OCT)

3.2.1. Background

Ophthalmic examination has been widely recognized as an effective method for the early diagnosis of systemic diseases and ocular disorders that may lead to severe vision impairment or blindness. Among various imaging modalities, Optical Coherence Tomography (OCT) has emerged as a powerful tool, providing high-resolution cross-sectional and volumetric images of the retina. This technique enables ophthalmologists and other specialists to visualize the three-dimensional structure of retinal layers in a non-invasive manner.

The retina itself is composed of multiple distinct layers, each with specific physiological functions. Quantitative information derived from these layers—such as retinal thickness or inter-layer boundary positions—serves as an objective biomarker for disease progression in conditions such as glaucoma, diabetic retinopathy, and age-related macular degeneration. Accurate delineation of these boundaries, therefore, plays a crucial role in clinical decision-making, longitudinal disease monitoring, and therapeutic outcome evaluation.

However, the segmentation of retinal layers from OCT images is a non-trivial task. Retinal structures exhibit high inter-patient variability, and disease-induced deformations often obscure the regular layered appearance. Moreover, OCT images are inherently affected by speckle noise, low contrast in deeper regions, and motion artifacts, which further complicate automated segmentation. Figure 45 shows an example of OCT and colored lines are boundaries of each retinal layer. You can see that it is difficult to determine the boundaries due to its image quality and disease cases are much more difficult to do segmentation. These factors demand robust algorithms that can adapt to both normal and pathological variations in retinal morphology.

Over the past two decades, numerous methods have been proposed to address this problem. Early approaches often relied on edge detection, graph-based models, or deformable contour techniques to extract layer boundaries. While these methods demonstrated promising results in normal retinas, their performance frequently degraded under pathological conditions.

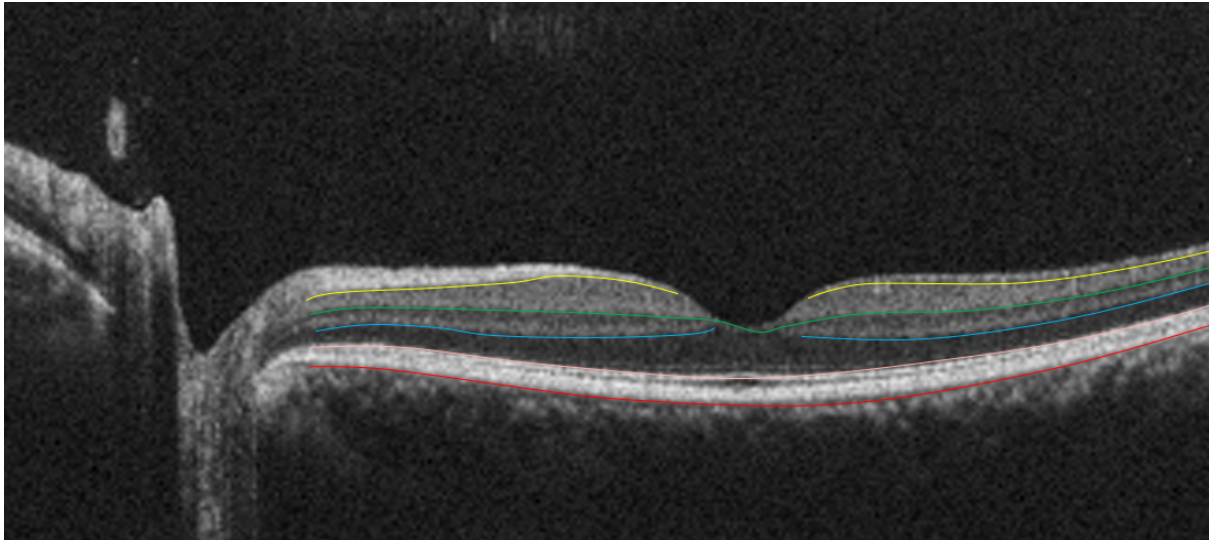


Figure 45 An example of OCT and boundaries of retinal layers.

This image is from <https://github.com/Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET?tab=readme-ov-file>

3.2.2. Proposal method

As this is research and development conducted at a private research institute, I cannot disclose details, however I will briefly explain what I did based on the patents or academic papers I applied (because they are disclosed in public).

We performed segmentation of the retinal layers using depth profiles. First, we filtered the OCT to remove noise and highlight the boundaries. Then, we created profiles as shown in the figure. After that, we classify the profiles by classifier that considers intensity information between layers. Finally, we can identify each peak in the profiles corresponds to the boundary of which layers.

Based on this strategy, we improved the algorithm in order to apply to various cases, for instance, the case that an artifact caused by blood vessel exists such as Figure, unclear boundaries due to retinal disease, etc.

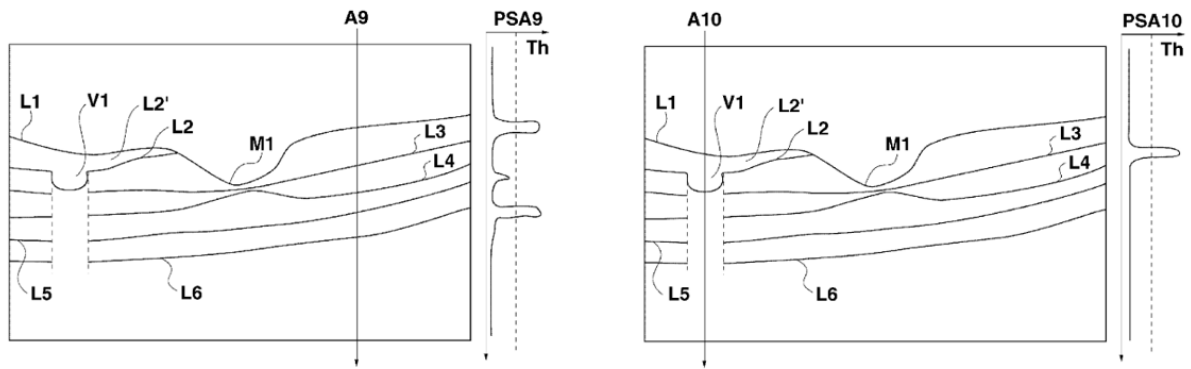


Figure 46 Profiles made by intensity information on OCT
 These figures are from patent "US9398846B2" in which one of the inventors is the author.
<https://patents.google.com/patent/US9398846B2/en>

3.3. Organ segmentation in CT

3.3.1 Background

Computed Tomography (CT) has become an indispensable modality in the diagnosis and management of abdominal diseases, including liver and spleen disorders. Accurate delineation of these organs is crucial for a wide range of clinical applications such as tumor detection, treatment planning, and postoperative evaluation. Consequently, the development of robust and precise segmentation techniques for liver and spleen in CT images remains an important research topic in medical image analysis.

Despite its clinical importance, automatic segmentation of abdominal organs in CT images is still challenging. One major difficulty arises from the variability introduced by the use of contrast agents. Depending on the time elapsed after injection, CT scans are acquired in different phases, typically classified into non-contrast, arterial, portal venous, and delayed phases. The gray-level distribution of the same anatomical structure can vary significantly across these phases, making it difficult to design a single image processing parameter set that performs reliably in all cases. Furthermore, the anatomical boundaries between adjacent organs (e.g., liver and stomach, spleen and kidney) are often ambiguous, further complicating accurate segmentation.

Traditional approaches have employed probabilistic atlas-based methods combined with Maximum a Posteriori (MAP) estimation to address this problem. For example, Park et al. constructed an abdominal probabilistic atlas and demonstrated its use in liver and kidney segmentation[40]. This approach relies on prior distributions of gray levels for each organ; however, since these distributions are strongly phase-dependent, accurate segmentation becomes infeasible when phase information is missing or incorrectly recorded. As a result, such methods suffer from decreased robustness in clinical practice.

Motivated by these challenges, this study aims to develop a robust segmentation method for liver and spleen that can cope with variability in contrast-enhanced CT images and provide stable performance across different phases.

3.3.2 CT phase estimation

Building upon the challenges discussed in the preceding sections—particularly the phase-dependent variability of CT image intensity distributions and the resultant instability in conventional atlas-based segmentation—this study adapts and formalizes the technical innovations described in US 10672111 B2 into a research-oriented framework. The key contributions of this work are summarized as follows:

Phase-Robust Two-Stage Segmentation Strategy

Inspired by the patented system, our method employs a two-stage segmentation approach:

- First, it extracts a primary anatomical region that exhibits minimal positional variation across subjects (e.g., liver or heart), doing so without relying on phase information—thus ensuring robustness when such information is missing or unreliable.
- Second, it uses features derived from this region to estimate the imaging phase, which then informs the segmentation of secondary target organs (e.g., spleen, kidneys) with greater accuracy by selecting phase-appropriate gray-level priors.

This two-stage paradigm enhances segmentation stability and precision in the presence of phase uncertainty.

Phase Estimation via Intrinsic Image Features

The method computes gray-level distribution features from the first extracted region to estimate the contrast phase (non-contrast, early, portal, delayed). This enables adaptive selection of segmentation parameters even when the image metadata is incomplete or incorrect, thereby overcoming a significant limitation of prior atlas-based methods.

Probabilistic Atlas Registration and Alignment Using Stable Landmarks

The invention employs anatomical landmarks (e.g., lung and pelvic centroids) to align a probabilistic atlas with the input image. By grounding this alignment in reliable, phase-invariant anatomy such as the points shown in Figure 47, our method achieves accurate atlas registration irrespective of contrast-induced intensity variations.

Phase-Specific Prior Modeling for Secondary Organs

Once the phase is estimated, the system selects or constructs organ-specific prior gray-level distributions tailored to that imaging phase, significantly improving segmentation accuracy for secondary organs like the spleen. When the phase is unknown or ambiguous, it can fall back on an averaged or portal-phase prior, enhancing robustness.

By translating these technical insights into our research context, we deliver a segmentation framework that:

- Adapts dynamically to contrast variability, addressing a core challenge in abdominal CT segmentation.
- Minimizes reliance on metadata, thereby enhancing applicability across heterogeneous clinical datasets.
- Combines atlas-based priors with phase-informed parameter selection, improving performance of precise segmentation of challenging organs such as the spleen.

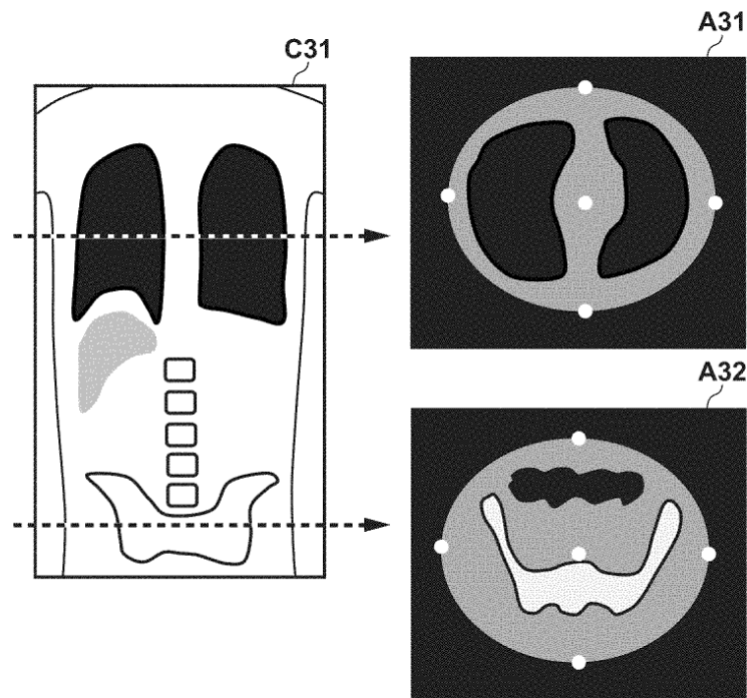


Figure 47 Phase invariant points in CT for image registration.
 These figures are from patent "US9398846B2" in which one of the inventors is the author.
<https://patents.google.com/patent/US10672111B2/en>

3.3.3 Refinement using Graph cut

To further improve the segmentation accuracy after phase estimation and atlas-based initialization, this study integrates a **graph cut–based energy minimization framework**.

Graph cut formulates the segmentation as the minimization of an energy function $E(L)$, where $L = \{l_p\}$ denotes the label of each voxel p as either belonging to the target organ or the background. The energy function is typically expressed as:

$$E(L) = \sum_{p \in P} D_p(l_p) + \lambda \sum_{(p,q) \in N} V_{p,q}(l_p, l_q) \quad (4)$$

Here:

- P the set of voxels,
- $D_p(l_p)$ is the data term, representing the cost of assigning label l_p to voxel p . In our case, this is derived from the likelihood estimated using phase-adaptive gray-level priors and atlas registration,
- $V_{p,q}(l_p, l_q)$ is the smoothness term, enforcing spatial consistency by penalizing neighboring voxels p, q that are assigned different labels,
- λ is a weighting parameter controlling the balance between fidelity to the data and smoothness of the segmentation boundary.

The data term is defined as:

$$D_p(l_p) = -\log P(I_p | l_p, \theta_{phase}) \quad (5)$$

where I_p is the observed intensity at voxel p , and θ_{phase} denotes the intensity distribution parameters estimated for the given contrast phase.

The smoothness term is commonly modeled as:

$$V_{p,q}(l_p, l_q) = \begin{cases} 0, & \text{if } l_p = l_q \\ \exp\left(-\frac{(l_p - l_q)^2}{2\sigma^2}\right), & \text{if } l_p \neq l_q \end{cases} \quad (6)$$

Where σ controls sensitivity to local intensity differences. This formulation preserves strong edges while promoting homogeneous regions.

By solving this optimization using the max-flow/min-cut algorithm, the method achieves an anatomically consistent segmentation that is robust against local intensity inhomogeneities, noise, and phase-dependent variability. In particular, this refinement significantly improves boundary delineation between adjacent organs such as the liver and spleen.

As explained at the beginning, due to restrictions on protecting confidentiality, we are unable to show the actual results images, however, I hope you understand this point.

3.6. Lymph node segmentation in CT

3.4.1 Background

Lymph nodes are one of the most important biomarkers in cancer staging and treatment planning. Detection and segmentation of lymph nodes from CT images are essential for tumor staging as well as for defining target volumes in radiotherapy. However, lymph nodes in CT scans are typically small, often only a few millimeters in diameter, and exhibit low contrast against surrounding tissues, making automated processing highly challenging. In addition, lymph nodes display great variability in shape, ranging from nearly spherical to irregular forms, which makes robust and generalizable detection difficult using traditional approaches.

Challenges

- Variability in size and shape: Lymph nodes vary widely, from very small nodes to enlarged, irregular structures.
- Ambiguous boundaries: Attenuation values in CT are often similar between lymph nodes, fat, and vessels, making boundary delineation difficult.
- Noise and artifacts: Small lymph nodes are often obscured by partial volume effects and image noise.

Previous Studies

Before the emergence of deep learning, research on lymph node segmentation primarily relied on atlas-based approaches, classical machine learning, and variational methods:

- Atlas-based segmentation: Slimane et al. proposed atlas-based registration to localize lymph node candidates [Slimane et al., 2009].
- Machine learning (SVM-based detection): Feulner et al. applied support vector machines (SVM) to CT image features for automated lymph node detection, showing improvements over threshold-based methods [Feulner et al., 2010].
- Graph cut / Level-set methods: Boykov and Jolly introduced an energy minimization framework using graph cuts for boundary segmentation in medical images [39], while Osher and Sethian's level-set method became a widely adopted tool for boundary tracking and lymph node/tumor segmentation [37].
- Shape models: Okada et al. proposed a statistical shape model-based approach for lymph node detection in CT, showing effectiveness especially for small lymph nodes [Okada et al., 2008].

Although these studies contributed important progress, challenges such as robustness against size variability, computational efficiency, and the need for accurate annotations remained unsolved.

3.4.2 Level set with scale and shape normalization

In order to solve issues in previous subsection, we integrate method with scale estimation using the Laplacian of Gaussian (LoG) with anisotropic scaling for shape normalization into level set method.

Scale Estimation using LoG

Lymph nodes are modeled as spherical “blobs,” and their representative size is estimated by finding the scale σ^* at which the LoG response is maximized.

$$\nabla^2 G(x, y, z; \sigma) = \frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2} + \frac{\partial^2 G}{\partial z^2} \quad (7)$$

where the Gaussian kernel is defined as:

$$G(x, y, z; \sigma) = \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\left(-\frac{x^2 + y^2 + z^2}{2\sigma^2}\right). \quad (8)$$

Shape Normalization via Anisotropic Scaling

Based on the estimated scale parameters, anisotropic scaling is applied along each axis to transform the target region into a nearly spherical form:

$$(x', y', z') = \left(\frac{x}{s_x}, \frac{y}{s_y}, \frac{z}{s_z}\right), \quad (9)$$

where (s_x, s_y, s_z) denote the scaling coefficients along each axis.

By using this methodology, we can convert the shape of various kind of lymph node to blob-like shape, such as shown in Figure.

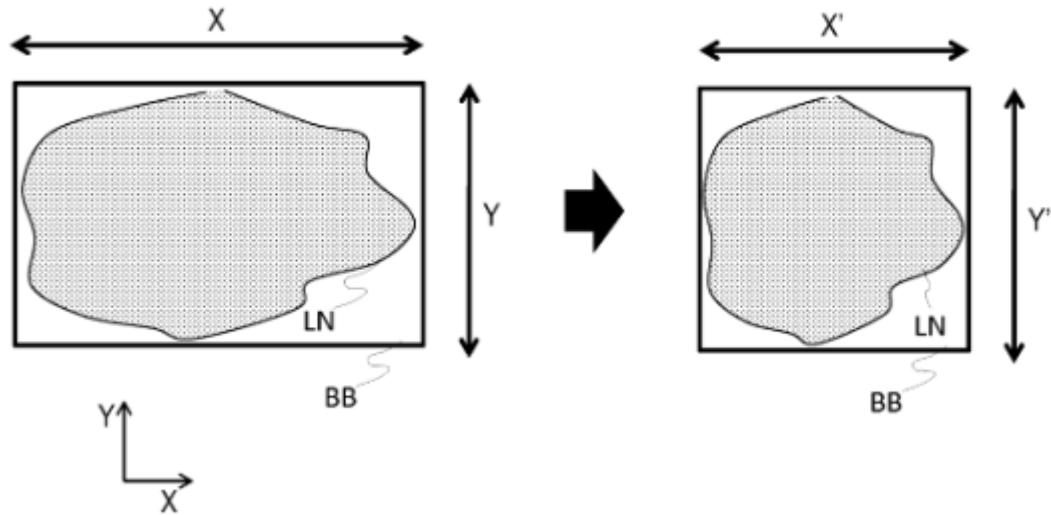


Figure 48 Blob shape conversion by using LoG.
 These figures are from patent "US9398846B2" in which one of the inventors is the author.
<https://patents.google.com/patent/US9741123B2/en>

Segmentation in Normalized Space

In this spherical space, segmentation can be performed more stably using graph cut [38] or the level-set method [37]. This normalization improves boundary detection robustness and reduces the dependency on lymph node size.

Back-transformation to Original Coordinates

Finally, the segmented region is transformed back to the original CT space, providing clinically applicable lymph node masks.

In summary of this methodology, we can insist the following contributions:

1. Automatic scale estimation via LoG enables consistent detection of lymph nodes across different sizes.
2. Anisotropic shape normalization improves the accuracy of classical segmentation methods (graph cut, level-set).
3. Integration of mathematical modeling and traditional methods extends the foundations of traditional approaches, overcoming their limitations and providing a novel hybrid framework.

3.6. Non-Alcoholic Steatohepatitis (NASH) assessment based on texture analysis of liver regions in CT images

From this subsection, I moved to the company named MEDIAN Technologies which locates in Sophia Antipolis in France. I worked here as a lead scientist for developing image processing and data mining algorithm. I will show what I researched in this company, however, same as the section before, contents are limited due to confidentiality. Therefore, I will try to explain only by using information disclosed in public.

3.5.1. Background

Non-alcoholic fatty liver disease (NAFLD) has become one of the most prevalent chronic liver diseases worldwide, closely linked to the increasing incidence of obesity, diabetes, and metabolic syndrome. A subset of NAFLD patients develop non-alcoholic steatohepatitis (NASH), characterized by hepatic inflammation, hepatocyte ballooning, and fibrosis. NASH is not only a progressive liver disorder but also a key risk factor for cirrhosis and hepatocellular carcinoma (HCC), which is the most common primary liver cancer and a leading cause of cancer-related mortality.

In the United States, NASH has already become the second leading cause of HCC and is projected to surpass viral hepatitis as the major etiological factor in the near future. The burden of NASH is increasing globally, placing significant strain on healthcare systems. Despite this, reliable methods for monitoring disease progression remain limited.

Traditionally, the gold standard for assessing NASH severity has been histopathological evaluation from liver biopsy. Biopsy allows assessment of steatosis, inflammation, ballooning, and fibrosis according to established scoring systems (e.g., NAS, SAF). However, biopsy has several limitations: it is invasive, associated with patient discomfort and risk of complications, and subject to sampling error due to the heterogeneous distribution of disease within the liver. Additionally, inter- and intra-observer variability in histological grading undermines reproducibility and reliability.

Given these limitations, there is a strong demand for non-invasive, quantitative, and reproducible biomarkers to assess disease progression and treatment response. Imaging modalities such as CT, MRI, and ultrasound have been increasingly explored for this purpose. Recent advances in quantitative image analysis and radiomics offer opportunities to derive imaging biomarkers that can objectively capture tissue heterogeneity and disease-related changes at the microstructural level.

3.5.2. Issues and challenges

Despite the progress in imaging technologies, several challenges remain:

1. Invasiveness and limited repeatability of biopsy
Biopsies cannot be performed frequently in clinical trials or routine practice, limiting their utility in longitudinal monitoring.
2. Subjectivity of histological scoring
Pathological evaluation of NASH is influenced by inter-observer variability, and small sampling areas may not represent the overall disease status of the liver.
3. Lack of standardized imaging biomarkers
Although conventional imaging modalities can detect advanced fibrosis or cirrhosis, there is no universally accepted imaging biomarker capable of tracking NASH progression across its different stages.
4. Disease heterogeneity
Both NASH and HCC are highly heterogeneous diseases. Methods are needed that can quantify this heterogeneity and provide objective measures that correlate with clinical outcomes.

These challenges highlight the need for advanced quantitative imaging techniques that can provide robust, reproducible, and non-invasive biomarkers.

3.5.3. GLCM based quantitative biomarker

The technical contribution of this study lies in the development of a non-invasive imaging biomarker based on quantitative texture analysis. In particular, the approach combines Gray-Level Co-occurrence Matrix (GLCM) for feature extraction with Earth Mover's Distance (EMD) for distribution comparison, providing a novel framework for liver disease progression assessment.

First, texture features derived from GLCM enable the quantification of subtle microstructural changes in liver tissue. As fibrosis progresses, tissue heterogeneity increases, leading to changes in local contrast and complexity observable in medical images. GLCM-based metrics objectively capture these disease-related alterations. Figure 49 shows schematic explanation how to compute textural signature and several examples.

In addition, this study employed EMD to measure differences between texture feature distributions across different disease stages. EMD is a distribution distance metric that quantifies the minimum "cost" required to transform one distribution into another. By applying EMD to GLCM-derived feature distributions, it became possible to evaluate disease-related differences that cannot be detected through simple mean-value comparisons.

Furthermore, the proposed approach was aligned with the clinically established METAVIR scoring system (F0–F4), which grades liver fibrosis from F0 (no fibrosis) to F4 (cirrhosis). By linking GLCM and EMD-based quantitative indicators with METAVIR scores, this framework

aims to establish clinically meaningful non-invasive biomarkers that reflect histopathological staging.

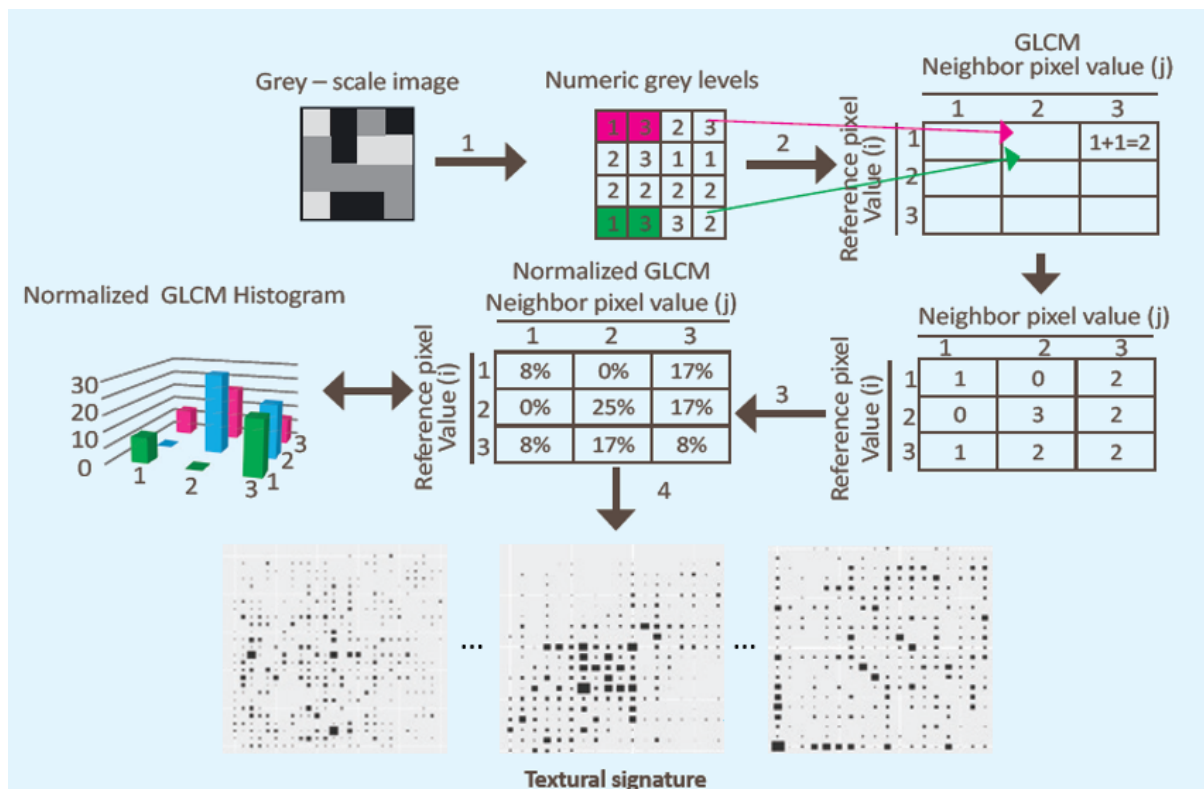


Figure 49 How to compute textural signature.
These figures are from article disclosed in the following line.
[MED_NASH_US LETTER_V1.indd](#)

3.5.4. Experimental results

Preliminary findings indicated that GLCM and EMD-based metrics demonstrated promising associations with fibrosis stages defined by the METAVIR score. Figure 50 shows examples of experimental results of METAVIR score classification by using our method. The key results are summarized as follows:

1. **Stage differentiation:** Texture features derived from GLCM showed clear differences between early stages (F0–F1) and advanced stages (F3–F4). Features related to contrast and heterogeneity were particularly sensitive to fibrosis progression and contributed to the discrimination between early and advanced disease.
2. **Effectiveness of EMD-based distribution comparison:** By applying EMD, the distance between feature distributions across METAVIR stages could be quantified. This enabled statistical differentiation even in intermediate stages (e.g., F2), where classification using single features is typically challenging.

3. Clinical relevance: By aligning GLCM and EMD-derived metrics with the METAVIR scoring system, the results suggested that fibrosis progression can be estimated without relying solely on histopathology. This highlights the potential of the proposed framework as a reproducible and non-invasive alternative for clinical practice.

In summary, this study demonstrates that the integration of GLCM-based features and EMD distribution analysis provides a novel quantitative imaging biomarker that reflects the METAVIR score. This framework holds significant promise for clinical trials, treatment monitoring, and future applications in therapeutic development.

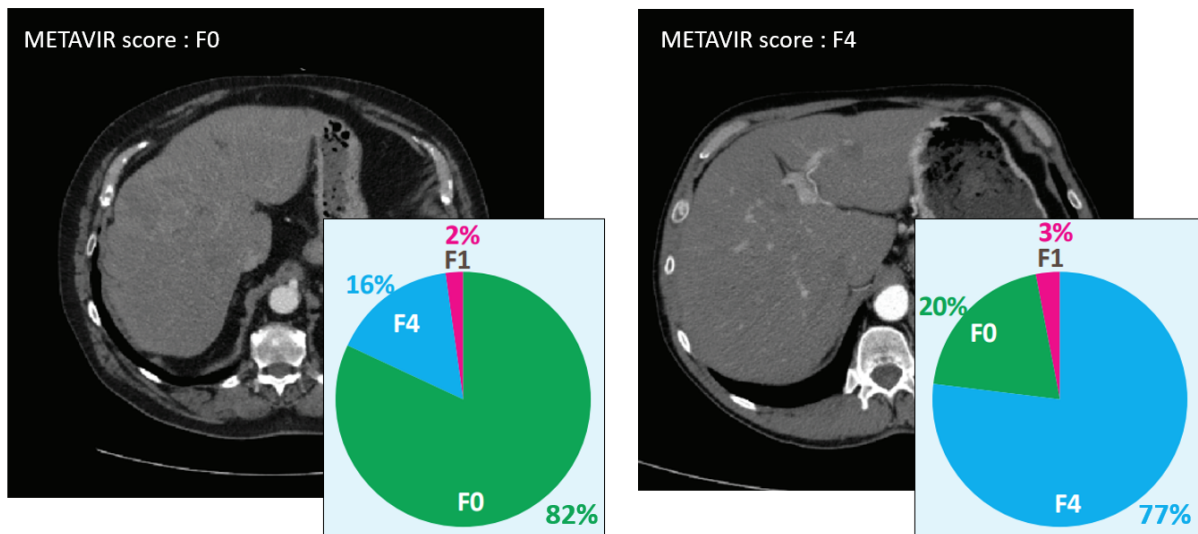


Figure 50 METAVIR score classification by using textural signature in the case of F0 and F4.
 These figures are from article disclosed in the following line.
[MED_NASH_US LETTER_V1.indd](#)

3.5.5. Conclusion

This chapter presented the development and application of normalized GLCM-based imaging biomarkers for assessing disease progression in NASH and HCC. The contributions can be summarized as follows:

1. A non-invasive and reproducible imaging method was proposed, overcoming the limitations of biopsy.
2. Texture-based analysis was demonstrated to capture subtle structural changes associated with disease progression.
3. Quantitative imaging biomarkers were introduced as potential tools for both clinical research and routine monitoring.

By reducing subjectivity and increasing reproducibility, this framework represents a promising step toward standardized non-invasive evaluation of liver disease. In the long term, such biomarkers could facilitate early detection, risk stratification, treatment monitoring, and therapeutic development for patients with NASH and HCC.

Future research should focus on validating these biomarkers in large multi-center cohorts, integrating them with other non-invasive methods (e.g., elastography, serum biomarkers), and exploring their predictive value for clinical outcomes such as cirrhosis progression and HCC incidence.

3.6. Benign/malignant classification of lung nodules using deep learning

The activities in this subchapter 3.6 are about lung nodule screening using Deep Neural Network for benign/malignant classification [41]. They were also conducted in MEDIAN Technologies. Due to confidentiality same as previous sections, its descriptions are limited, especially the concrete algorithm I developed and results cannot be shown. I hope you understand this point.

3.6.1. Background

Lung cancer remains one of the leading causes of cancer-related mortality worldwide. Early detection and treatment are crucial, yet in many cases the disease is diagnosed at an advanced stage, limiting therapeutic options and reducing survival rates. Computer-assisted detection (CAD) has emerged as a promising technology to support early identification of pulmonary nodules, thereby improving diagnostic accuracy and efficiency in clinical practice.

The National Lung Screening Trial (NLST) demonstrated the efficacy of low-dose computed tomography (CT) in the early diagnosis of lung cancer. The NLST database provides a unique resource that includes a large collection of CT images together with structured clinical and outcome data. Leveraging this dataset, the present research aims to enhance the accuracy of CAD systems for pulmonary nodule detection and malignancy assessment [42].

With the increasing availability of deep learning techniques, there is strong interest in developing algorithms that can reduce false positives and improve detection sensitivity. Nevertheless, despite encouraging research results, the robustness and reliability of these systems in real-world clinical environments remain limited. Therefore, careful validation is required before widespread clinical deployment [43][44].

3.6.2. Issues and challenges

In order to make this technology practical in clinical situation, we can list the following challenging points.

1. High False Positive Rate and Reliability: Current CAD systems for pulmonary nodule detection tend to produce numerous false positives, which hinder their clinical adoption. Excessive false alerts may lead physicians to recommend unnecessary follow-up imaging or invasive procedures, increasing patient burden and healthcare costs. Thus, developing

algorithms that maintain or improve sensitivity while significantly reducing false positives is an essential challenge [45].

2. **Malignancy Assessment:** Beyond simple detection, CAD systems must provide information about the likelihood of malignancy. Accurate malignancy prediction is critical for determining appropriate clinical follow-up and management strategies. This requires advanced classification models that can differentiate between benign and malignant nodules with high confidence [46].
3. **Utilization of the NLST Dataset:** The NLST dataset, with its combination of CT imaging, annotations, and patient outcome information, represents an invaluable training and validation resource for algorithm development. However, working with such a large and heterogeneous dataset raises challenges including data preprocessing, bias control, label accuracy, and ensuring model generalizability [42].
4. **Balancing Performance and Robustness for Clinical Application:** For clinical implementation, CAD systems must not only achieve high sensitivity and specificity but also demonstrate robustness across diverse imaging conditions and patient populations. In addition, issues of reliability, accountability, and potential legal implications must be addressed before integration into routine clinical workflows [43][47].

3.6.3. Technical contributions

As explained at the beginning of this subsection, the details of technical parts and results cannot be mentioned. However, at least we can say that, we attempted to estimate the malignancy of nodules using deep learning. In particular, we estimated the likelihood of benign nodules becoming malignant in the future using deep learning, with the aim of detecting the timing of the transition from benign to malignant as early as possible by observing changes over time in the same patients.

3.7. Summary

The researches in this chapters were my first several works after obtaining PhD. Through this work, I learned a lot especially about how to make a robust system, the importance of validation because the technologies we developed might be installed on real product. In this epoque, although I did not use cutting edge methods so often. However, I combine some traditional methods with mathematical models in order to solve the issues that we were facing. Even though I did not use recent methods, at least I thought they are robust solutions for each purpose. As a result, I could apply many patents in this period.

Chapter 4

Sensor data processing with AI in Automotive domain

4.1. Introduction

After working in medical domain over 10 years described in Chapter 2 and 3, I moved my research field to automotive domain. To be more specific, I joined IMRA EUROPE since 2018, who is a research institute and is a branch of AISIN corporation. As AISIN corporation is one of the biggest car parts suppliers, we are doing research for instance to improve performance of powertrain control, sensor processing, etc. Especially, computer science department in IMRA EUROPE, where I'm belonging to, is in charge of object detection/recognition on various sensor data using AI such as deep neural network.

Since joining this company, I have been involved in various projects, for instance, object detection on LiDAR (3D point cloud) data, data analysis on production line in factory, AI motor control and so on. In this chapter, I explain the following three projects in which I participated as both roles of manager and developer:

- Radar (milli-wave) signal processing for passenger detection in vehicle
- Pedestrian detection on event camera (Dynamic Vision Sensor (DVS)) data
- Feasibility study of ANN2SNN conversion (SNN: Spiking Neural Network)

In all projects, I used deep neural network. However, we did not use RGB camera which is the most popular in these kind of perception themes because we tried to develop systems that work with low energy consumption. In fact, about milli-wave and event sensors, the sensor itself consumes less energy than RGB sensor. In contrast, it is more difficult to detect objects on their data because the data include more noise or less visual information. Then, from the algorithmic point of view, SNN also has a big potential to reduce energy consumption compared with conventional deep neural network (DNN) such as convolutional neural network (CNN) due to less signal propagation. However, same as the sensors, there are some difficulties to reach the same level of performance as the conventional DNN with rich signal propagation.

The subsections below show how did I solve these issues in each topics. In this chapter also, as all activities were done in private research institute, disclosure of their contents are limited because of confidentiality. Therefore, I explain by using publications and patents which have been already disclosed in public.

4.2. Radar signal processing for passenger detection in vehicle

4.2.1. Background

As vehicle intelligence evolves, various sensors are being installed for tasks like object detection and self-localization. Among these, radar (millimeter-wave) sensors are particularly popular due to their low price and easy installation. We are also developing a detection system using millimeter-wave sensors, aiming not only to reduce costs by replacing older sensors but also to enable new applications leveraging the sensor's unique characteristics.

However, millimeter-wave signals are susceptible to noise, including reflected waves, which can degrade detection performance. Although signal processing techniques using Fourier transforms have been applied to generate distance maps and distinguish between meaningful data and noise, they may also unintentionally remove significant signals. This issue becomes especially critical in confined environments like a vehicle cabin, where differentiating signal types is more complex.

We hypothesize that the selection and processing of these signals such as the Fourier transform significantly affects detection performance. Therefore, we implemented a deep neural network that bypasses traditional preprocessing and directly uses raw sensor data, allowing the network to learn optimal signal selection. Specifically, the sensor's complex-valued signals are processed in a way that preserves the relationship between their real and imaginary components, maximizing information retention and enabling effective noise removal through training.

In the recent trend toward smarter vehicles, millimeter-wave sensors are increasingly used due to their ease of installation and low cost. However, the signals obtained by these sensors often contain a lot of noise, such as reflected waves, making it difficult to achieve high detection accuracy even with noise reduction techniques like Fourier transform. To address this issue, we implemented a deep neural network that directly uses raw sensor data, improving object detection performance. Especially, we applied implemented neural network to in-vehicle passenger detection.

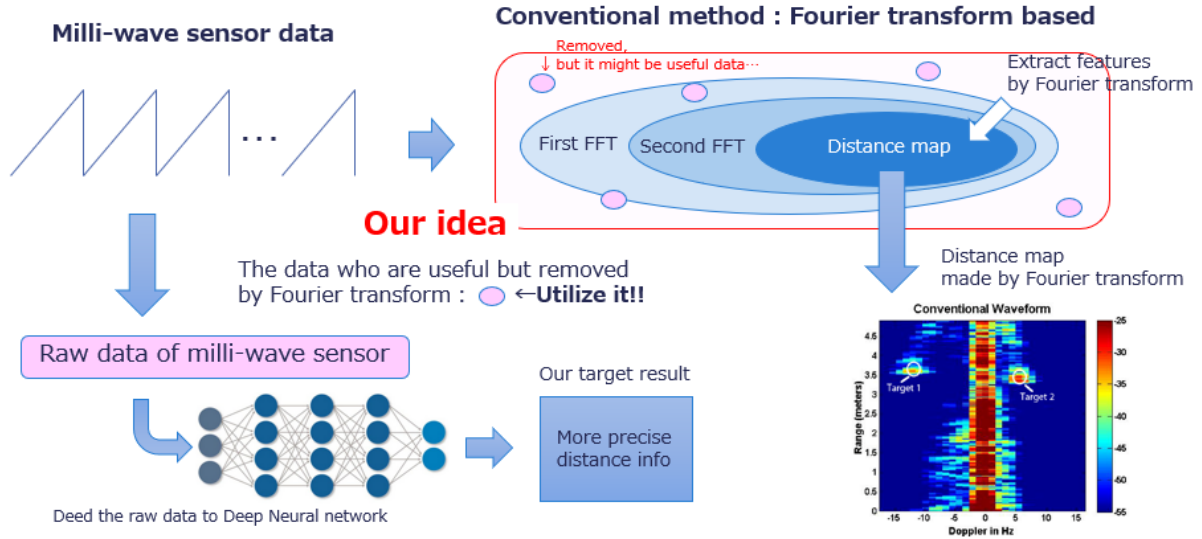


Figure 51 Conventional method and our idea explained in this section.

4.2.2. Conventional methods

The standard approach for object detection using millimeter-wave sensors involves applying multiple Fourier transforms to generate distance maps. This section introduces conventional noise removal techniques using filtering and deep learning methods that take distance maps as input.

Noise removal by filtering

One common approach is to use filters like band-pass filters to eliminate specific noise components. However, setting threshold values for valid frequency ranges requires expert knowledge and trial-and-error. While hand-tuned filtering may work in controlled environments, dynamic or complex scenarios (like object detection) often contain nonlinear noise beyond the scope of such simple methods.

Approaches using Deep Neural Networks

Millimeter-wave sensors have seen a dramatic rise in adoption within the automotive industry, particularly due to their resilience to adverse weather conditions and their cost-efficiency compared to LiDAR or camera-based systems. Traditional object detection methodologies that rely on Fourier transform techniques attempt to filter out noise and extract features from radar spectrograms. However, these approaches often rely on hand-crafted signal processing pipelines, which can limit generalizability and adaptability in dynamic environments such as vehicle interiors, where multiple reflectors, occlusions, and seating configurations introduce complexity.

After the epoch of the hand-crafted signal processing, a widely used modern approach involves generating distance maps through multiple Fourier transforms and inputting these into a neural network[48]-[50]. Its pipeline is shown in Figure 52. While the network can learn

to mitigate noise left over from the transforms, some useful signals might already be lost in the preprocessing phase. Thus, this method has an inherent limitation in performance due to this information loss, as explained in Figure 51.

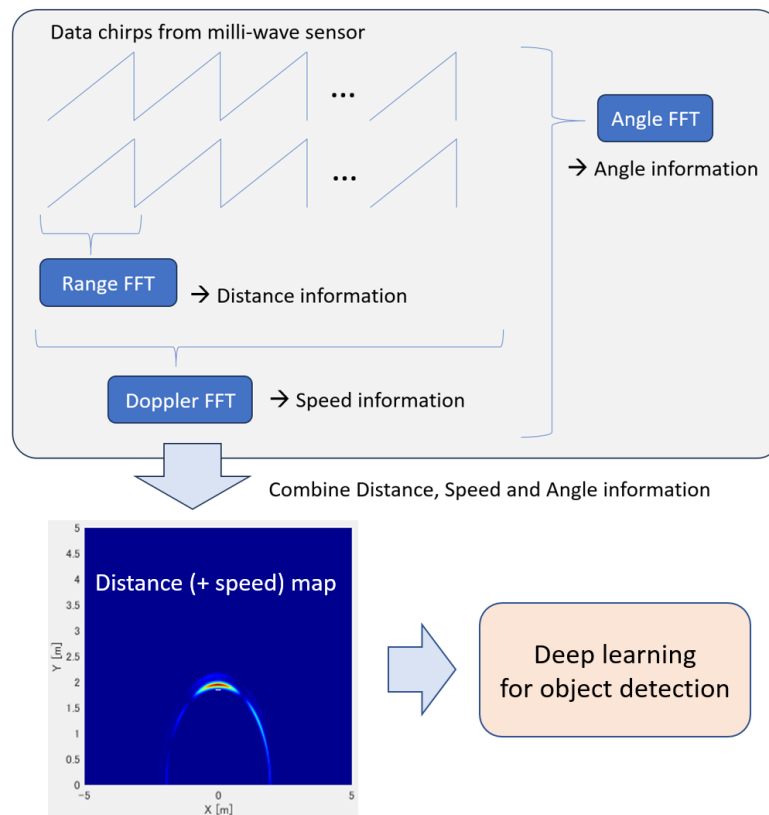


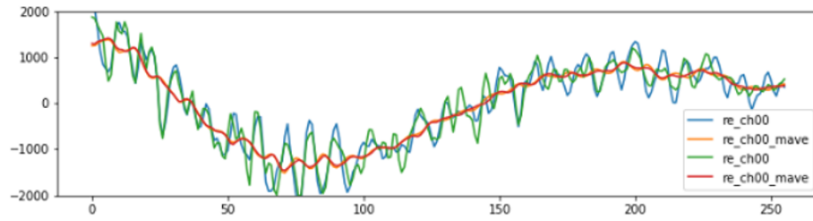
Figure 52 Pipeline of method using deep learning with distance map

4.2.3. Technical contributions

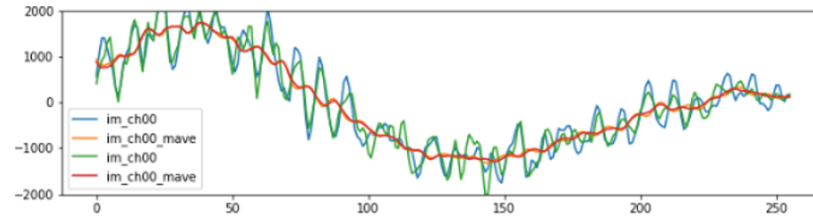
In our study, by bypassing the classical preprocessing stage and feeding raw complex-valued radar signals directly into a carefully designed deep neural network, we demonstrated significant improvements in seat occupancy detection tasks. The coupling of real and imaginary parts ensures that phase relationships critical for target localization and classification are retained. Our results indicate that such architecture can substantially outperform traditional systems, especially in challenging, cluttered environments like the vehicle cabin.

Input layer for complex-valued raw data

The raw data from our sensor consists of complex numbers (real and imaginary parts). In the input layer, separate channels are assigned to the real and imaginary parts, which are then processed in a coupled manner, preserving their correlation. In a preliminary comparison experiment with a network that treated the two parts independently, the coupled-input design showed higher accuracy.



Raw data + its moving average of real part value



Raw data + its moving average of imaginary part value

Figure 53 Raw milli-wave sensor data composed of complex numbers

Network architecture

The architecture includes an input layer designed for coupled real/imaginary values, three convolutional layers, two fully connected layers, and an output layer predicting occupancy probabilities for five seats (second and third row), outputting five values indicating presence/absence of occupants.

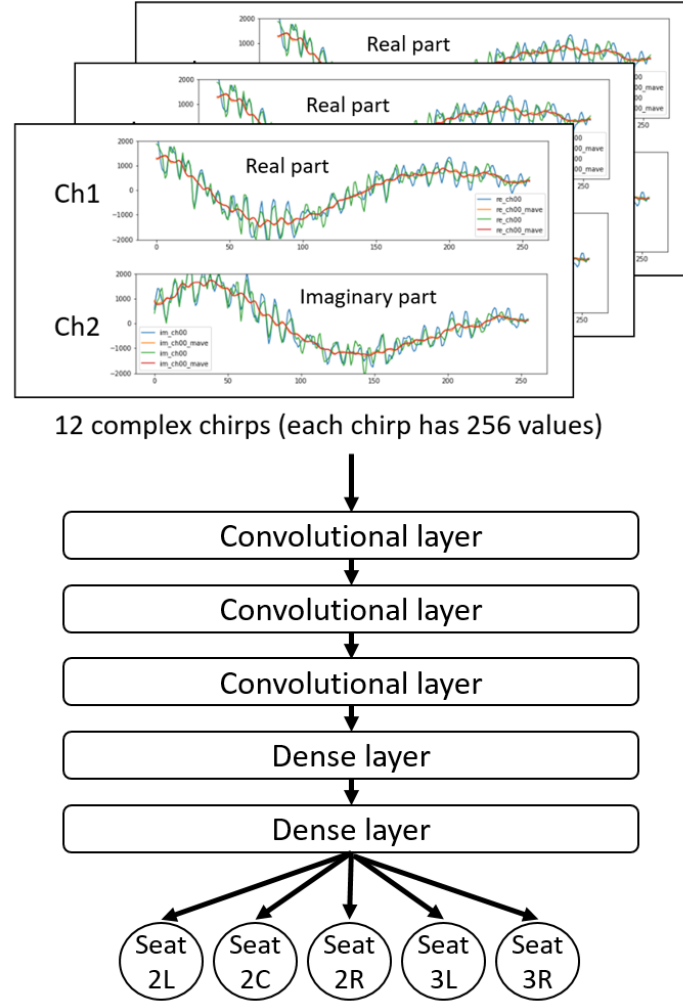


Figure 54 Architecture of our neural network which input raw milli-wave sensor data

Experiments

This section details the experimental setup and results comparing the conventional approach and our proposed method.

Comparison with conventional method

We implemented a network using Fourier transform-based preprocessing, replacing the input layer with a frequency-domain input. By comparing this with the raw-data-based network, we assessed the impact of the preprocessing on performance.

Experimental data

Data was collected with a millimeter-wave sensor mounted on the car ceiling, targeting five seats: three in the second row (2L, 2C, 2R) and two in the third row (3L, 3R). We captured presence/absence for each seat, generating 32 combinations through repeated trials with different subjects. Each pattern contains about 3000–9000 samples for training, whose distribution is shown in Figure 55. Then, test data, which is shown in Figure 56, includes about

1500 samples per pattern. The experiment was conducted on data from two different vehicle models (vehicle A and B).

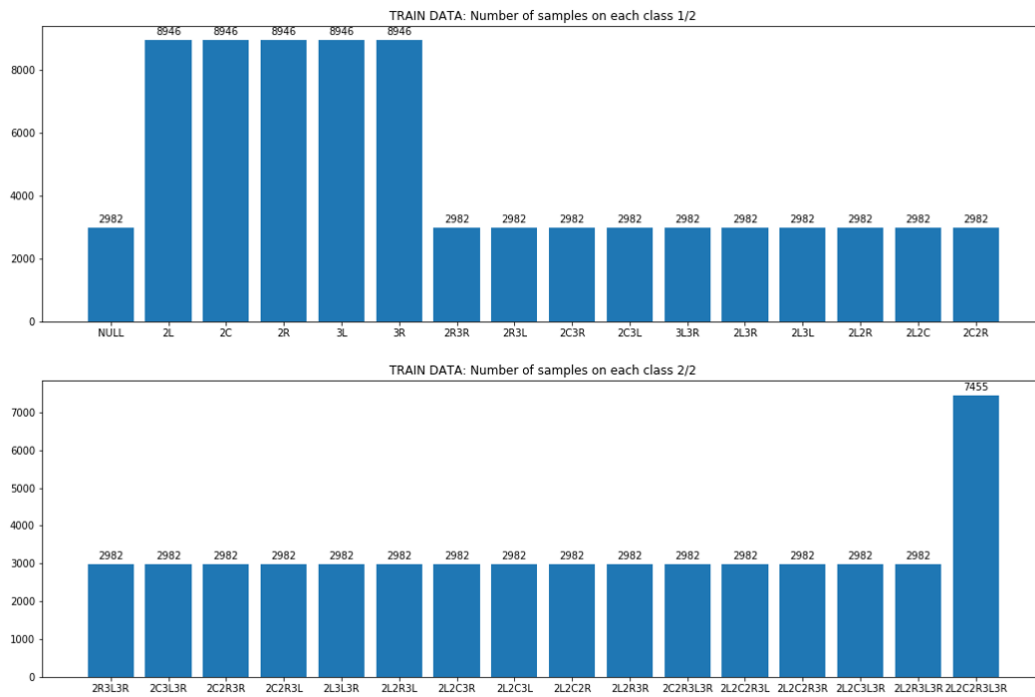


Figure 55 Data distribution in each class in training dataset

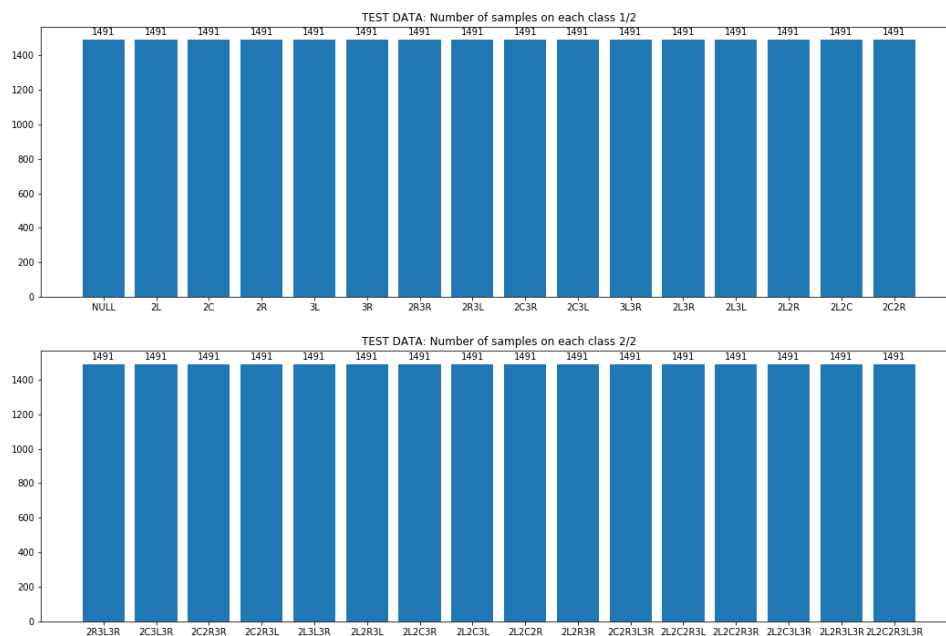


Figure 56 Data distribution in each class in testing dataset

Experimental results

Three methods were evaluated: 1) Fourier transform preprocessing + neural network (conventional method), 2) Raw data without coupling of real/imaginary parts, 3) Raw data with coupling (proposed method). Results showed that raw data input improved performance for both vehicle models compared to the Fourier method. Furthermore, coupling led to a 30.9% increase for the vehicle A and 17.6% increase for the vehicle B in detection accuracy.

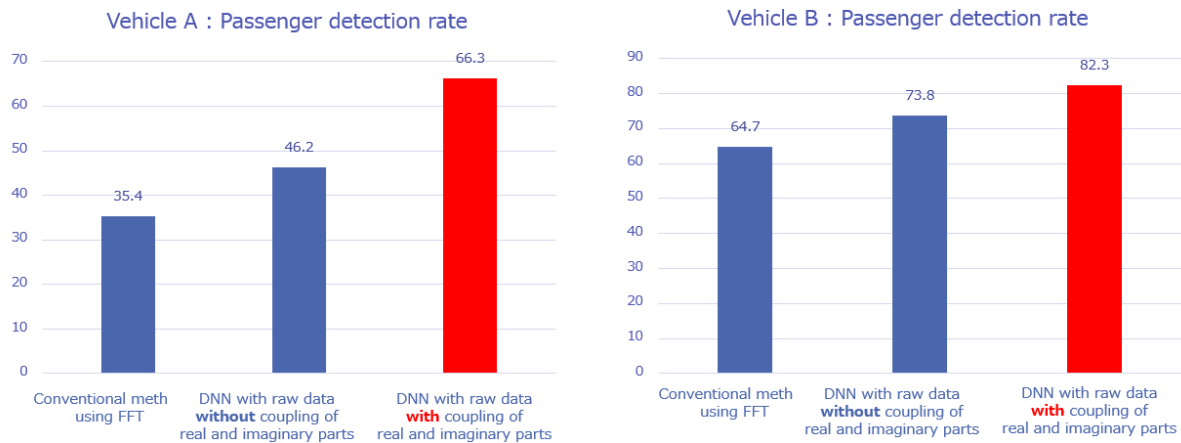


Figure 57 Passenger detection results in each method

Conclusion

We implemented a deep neural network that directly uses raw millimeter-wave data, avoiding traditional noise filtering, and verified its effectiveness. The results demonstrated a significant performance gain, showing the potential for fully leveraging millimeter-wave data.

Future work will involve optimizing the network considering sensor design and characteristics, further improving accuracy and testing robustness across more varied datasets with a view toward commercialization.

4.3. Pedestrian detection on event camera data

4.3.1. Background

With the recent trend of electrification in the automotive industry, it has become possible to utilize battery power for various devices and sensors, leading to numerous proposed applications. However, the increased power consumption of these devices reduces the driving range of electric vehicles, thereby creating a pressing demand for low-power system development.

One promising solution is the event camera, also referred to as a Dynamic Vision Sensor (DVS) or Neuromorphic Vision Sensor (NVS). Inspired by the biological retina, this novel vision sensor fundamentally differs from conventional RGB cameras. Unlike traditional imagers that output frames of dense pixel values at fixed intervals regardless of luminance change, event cameras detect changes in luminance and output pulses only from pixels where events occur. Pixels without changes remain silent. As a result, the acquired data consists of sparse time-series events, each containing coordinates, polarity, and timestamps. From the viewpoint of power consumption, this feature enables significant energy savings, particularly in environments with limited luminance variations.

Nevertheless, the data obtained from event cameras—hereafter referred to as event data—differs considerably from frame images captured by RGB cameras. Figure 58 shows several examples of images reconstructed from event data. This discrepancy complicates the application of conventional image-processing algorithms. For example, convolutional neural networks (CNNs), which are widely employed in object detection and recognition, require dense pixel-based frames as input. Event data, by contrast, provides binary polarity information (+/−) rather than graded intensity levels, often resulting in insufficient information for high recognition accuracy. Furthermore, event data tends to be noisy and may only capture regions of brightness change within objects or backgrounds. Consequently, objects often appear fragmented, leading to both false detections and missed detections. Addressing these issues has become a major research focus.

In this subsection, we report improvements aimed at enhancing the accuracy of object recognition systems based on event cameras. Specifically, we investigate the introduction of alternative event-based features beyond polarity for CNN input, and we propose improvements in the detection (bounding box generation) stage prior to recognition. For validation, we employ an in-house dataset designed for pedestrian detection around vehicles, and we conduct deep learning-based training and evaluation.

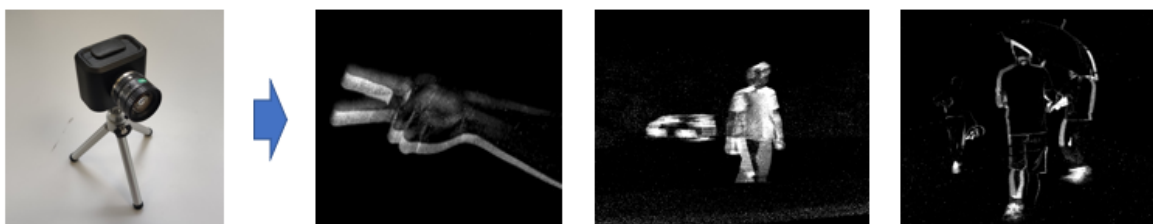


Figure 58 Examples of images reconstructed from event data

4.3.2. Conventional methods

We are currently developing a pedestrian recognition system using event cameras for surveillance purposes. In addition to their energy efficiency, event cameras provide a wide dynamic range (120 dB), enabling their use in diverse environments such as nighttime, indoor,

and in-vehicle scenarios. Moreover, their high temporal resolution (1 Mframe/s) allows advanced applications such as motion recognition and behavior estimation.

In preliminary studies, we successfully developed a gesture recognition system combining event cameras with deep learning. Compared with a system using RGB cameras and CNNs, the event-based system maintained comparable recognition accuracy while reducing power consumption by 96% [51][52]. Building upon these findings, we are extending our research to more general-purpose recognition systems. Unlike the controlled environment of gesture recognition, general pedestrian detection must address challenges such as background noise and variations in object size due to changes in camera distance.

Previous works have proposed several methods for object recognition with event data [53][54]. A common approach is to segment event streams into time windows to generate frame-like representations, mapping polarity values into pixel space for CNN input. For example, Singla et al. [55] generated binary images from event polarities, applied morphological filtering for noise removal, performed labeling and thresholding to detect bounding boxes, and finally cropped and classified the candidate regions using a CNN. This pipeline—first detection, then recognition—follows the canonical strategy of object recognition and achieves both high accuracy and low power consumption. Consequently, we adopt this method as a benchmark.

However, through comparative studies, we identified two limitations. First, in the detection stage, background noise caused by subtle movements or weather-induced brightness fluctuations significantly degrades performance. Although morphological filtering and labeling alleviate this issue, improper parameter tuning often leads to excessive false detections or missed objects. Second, in the recognition stage, relying solely on polarity information proves insufficient for handling more complex recognition tasks. Richer data representations leveraging additional event information are required.

4.3.3. Proposed method

In order to address these two challenges, we propose two improvements. For detection, we employ a patch-based approach, dividing frame images into grids and using event-based features within each patch to determine candidate regions. For recognition, we introduce alternative data representations for CNN input, specifically (i) timestamp-based intensity maps and (ii) event-count accumulation maps, either individually or in combination.

Patch-based detection

Conventional binary image processing with morphological filtering often leaves residual noise, causing frequent false detections. Moreover, since event data only registers luminance changes, even single objects may appear fragmented across multiple regions. Although end-to-end deep learning approaches incorporating bounding box detection within the network can improve accuracy, they impose high computational and power demands unsuitable for in-vehicle edge systems.

Therefore, we divide frame images into grid patches and detect candidate regions based on event counts within each patch. Patches exceeding a threshold are marked as detections. As demonstrated in Figure 59, this approach successfully suppresses false detections while preserving true object events.

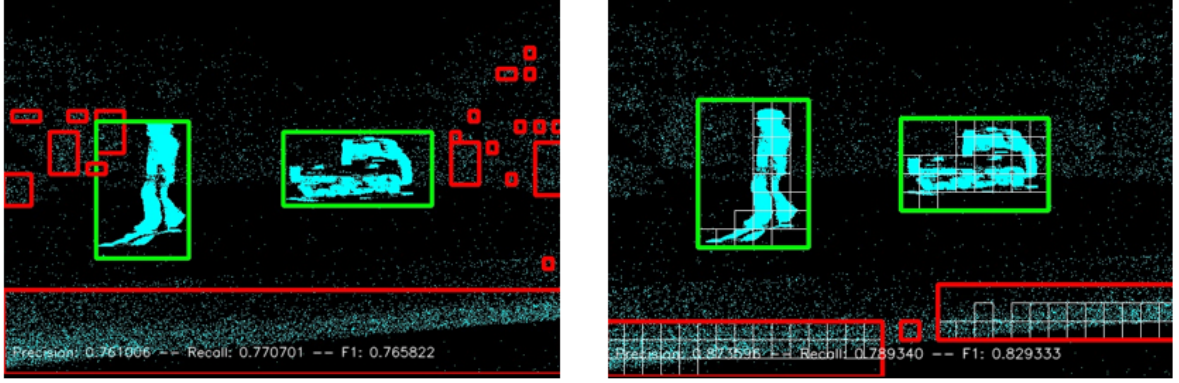


Figure 59 Examples of object detection on event data with and without using patches, left hand side figure is “without using patch”, right hand side is “with patches”

Data representation considering timestamp and event count

Each event contains a timestamp, which we normalize between the frame’s start and end times to assign grayscale values:

$$I_{ts} = G \cdot \frac{t - t_{start}}{(t_{end} - t_{start})} \quad (10)$$

where I denotes pixel intensity, t is the event timestamp, t_{start} and t_{end} are the frame’s temporal bounds, and G is the number of intensity levels (256 levels in this study).

Additionally, we generate images representing the accumulated number of events per pixel. Such data captures repeated activity at the same location due to vibrations or overlapping objects, offering valuable motion-related cues. Figure 3 illustrates examples of binary polarity images, timestamp maps, and event accumulation maps.

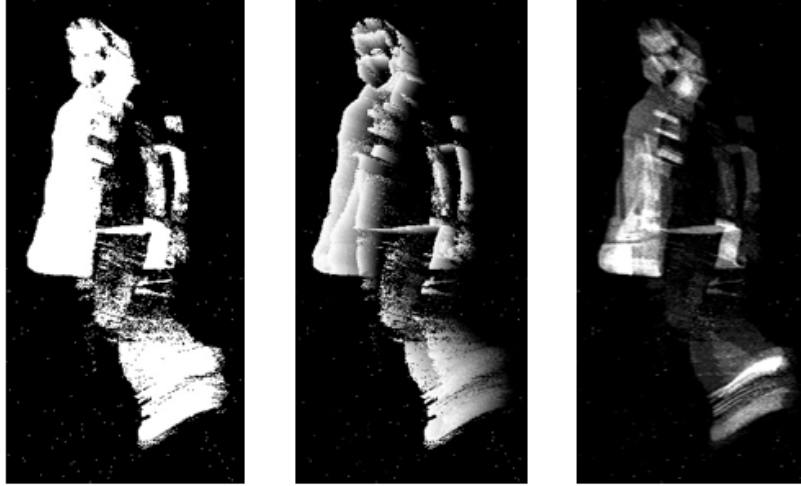


Figure 60 Data representation as image by using binary (left), timestamp (center) and count (right)

4.3.4. Experimental results

We applied the method described in the previous section to a dataset created for the purpose of recognising people outside vehicles and conducted experiments to compare its recognition accuracy with that of conventional methods.

Dataset

We constructed a dataset for binary classification (pedestrian vs. non-pedestrian), considering variations in subject count, camera distance, weather, and background. For training, we extracted 28,236 bounding boxes from 89 videos, allocating 90% for training and 10% for validation. For evaluation, 439 frame images were prepared for bounding box detection and recognition using the trained CNN.

Evaluation and comparison of performances

We conducted experiments using the above dataset. To evaluate the proposed method, we implemented the benchmark method HyNNA (Hybrid Neural Network Architecture) described in Section 4.3.2, and replaced the bounding box detection part and the data representation part for input to the CNN with our proposed method to observe the change in performance.

First, we compared the detection results obtained using the morphological filter and labelling process of the conventional method with those obtained using the patch segmentation method described in the section named “Patch-based detection”. The results are shown in Figure 61. As can be seen from the figure, our proposed method achieved better results in terms of F1 score compared to the conventional method. This is primarily due to the successful suppression of over-detection of background noise, as intended.

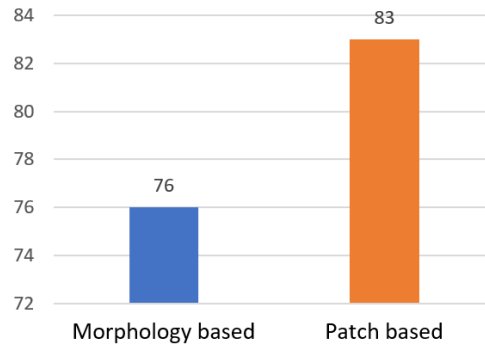


Figure 61 Detection performance “morphology based VS patch based”

Next, as a comparison experiment of data representation, we visualised the timestamps and event counts described in the section named “Data representation considering timestamp and event count”, input them into a CNN, trained the model, and then applied it to the evaluation data (this evaluation experiment did not include any improvements to the detection method using patch segmentation described above, and only evaluated the differences in data representation). The results of the conventional method and the proposed method are shown in Figure 62.

These results show that changing the CNN input from a binary image representation indicating only the presence or absence of events to a grayscale image representation indicating timestamps and the cumulative number of events significantly improved performance. Furthermore, combining these two approaches yielded even better results.

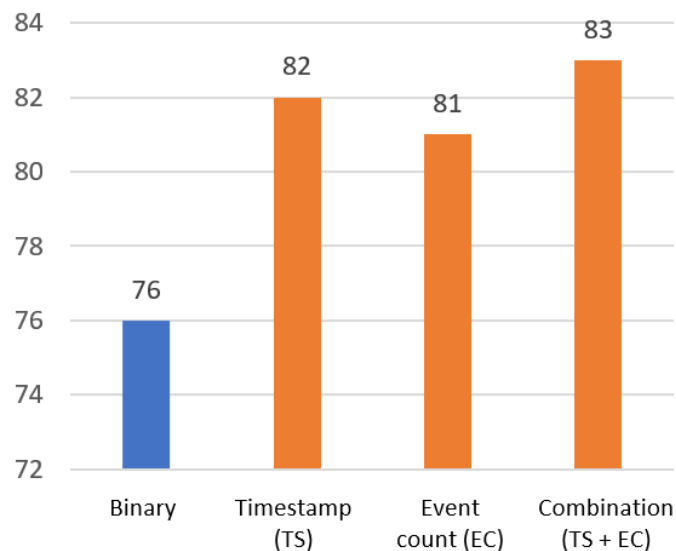


Figure 62 Detection performance between data representations

The reason for the improved performance is similar to that of the detection process: the number of false recognitions of background noise was reduced. In the conventional method, even background noise was often not only removed as noise during the detection process but also recognised as a person during the recognition process when the number of events was high. However, with the proposed method, providing more information to the CNN enabled the reduction of such misclassifications. As shown in the graph in Figure 63, the number of false positives (FP) in the recognition using timestamps is reduced to about one-third compared to the conventional method using binary images as data representation.

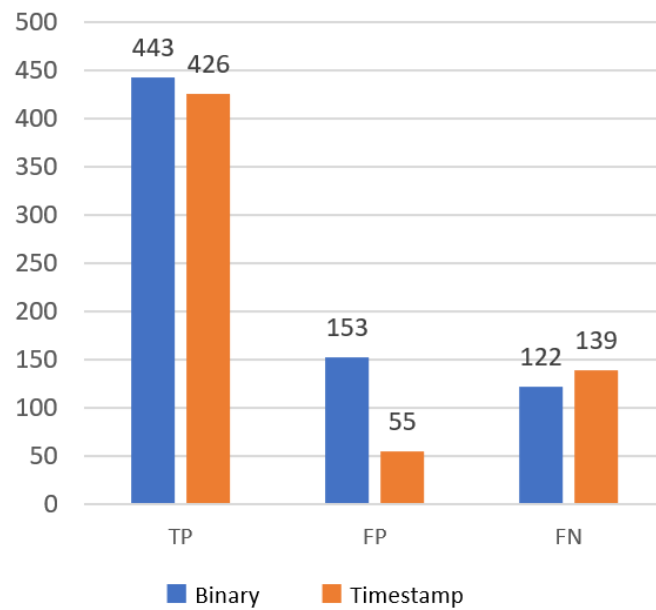


Figure 63 Number of FP (False Positive) in two data representations, Binary and Timestamp

Conclusion

This paper proposed two improvements for object recognition with event cameras: (i) patch-based bounding box detection and (ii) alternative data representations leveraging timestamps and event counts. Both approaches effectively reduced false positives and improved pedestrian recognition accuracy. Future work will focus on further enhancing recognition accuracy while simultaneously optimizing power consumption for real-world in-vehicle applications.

4.4. Feasibility study of ANN2SNN

4.4.1. Our mission and objectives

The application of Convolutional Neural Network (CNN) in various fields has recently been growing rapidly. However, CNN requires huge computation cost in compensation for its high

performance. Under this situation, Spiking Neural Network (SNN) draws attention because of its lower computation cost. As of today, several methods that convert from CNN to SNN have been disclosed. They have a potential to make SNN prevalent and practical technology.

As one of our mission is, to realize a vision system with super low energy consumption, which can contribute to sustainable development. That is why SNN approach is really attractive for us. We are confident that this technology will be one of the key technologies in automotive domain because there is now a big wave of electrification such as the increase of electric vehicle. Although there are several difficulties in SNN described in latter section, above CNN to SNN technologies is easy to handle because it solves one major issue, which is the difficulty of SNN training. Therefore, we investigated and tried to use this CNN to SNN technology as initial step.

After the preliminary study of CNN to SNN, we found that one drawback exists. The deeper architecture is, the more the performance decreases after conversion to SNN. In this report, we measured the impact of the SNN conversion, regarding the performance decrease as a function of CNN architecture. As a result of this investigation, we propose a way to identify optimal depth of architecture for SNN conversion. We think this result will be a reference for designing a neural architecture to be converted to SNN and ultimately applied in automotive.

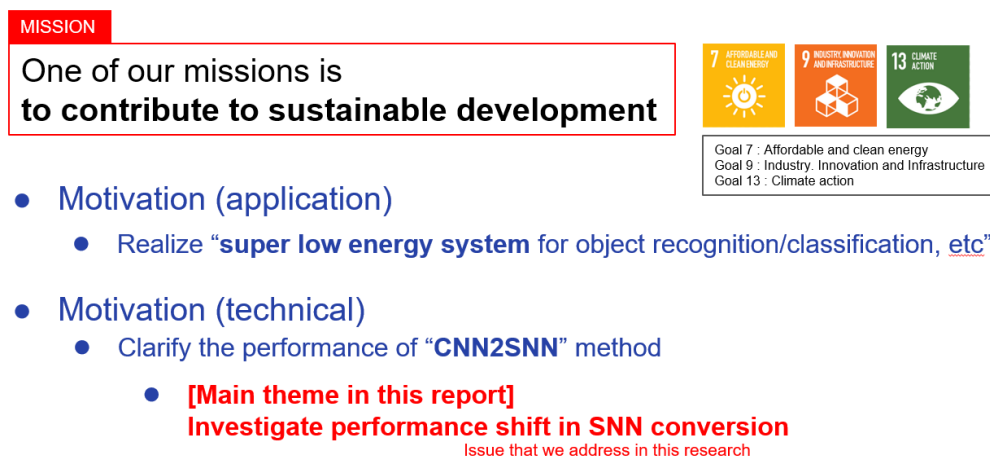


Figure 64 Our mission in this project and issue to which we addressed

4.4.2. Technical background

The objective of this project is to develop a vision system such as object detection/recognition with super low energy consumption. In order to struggle with this research topic, we investigated background from prior arts and clarified issues of current technologies. In this section, we explained technical background, issues and prior arts.

Computation cost in conventional methods

Everyone will agree that deep learning has achieved great successes in various fields such as object detection in image, speech recognition and text understanding. However, it is also

evident that one of the drawbacks of deep learning is its huge computational cost. It is a barrier for always-on systems or edge computing devices. Especially in automotive domain, we are always facing this issue when developing systems for as real-time object detection during driving and for surveillance system which require operating for long time while car is parked. In order to reduce computation cost, many researchers are focusing on optimizing the architecture of deep neural network (DNN) or developing dedicated hardware.

Spiking Neural Networks

As one of the solutions for this issue, the next generation of neural network is drawing lots of attention. It is the Spiking Neural Network (SNN) that is inspired by the behavior of human brain. To be more specific, SNN deals with asynchronous pulse (spike) signals that are derived when an event happens, such as shown in Figure 65. In other words, SNN does not need to do any signal propagation if there is no event. This methodology is completely different from conventional neural network, for example, Convolutional Neural Network (CNN) for image analysis. CNN must propagate all pixel values for every frame even if there is no change in frames. For this reason, we can easily understand that SNN will be able to realize an energy efficient system: propagating and processing only the changes. As of today, many publications and several libraries about SNN have been released [56][61].

SNN : Spiking Neural Network

- Inspired by human brain
- **“Spike” and “Asynchronized (sparse)” signal propagation**
 - Spike : easy operation (all calculations are done by “addition”)
 - Asynchronization : no signal propagation if it is unnecessary

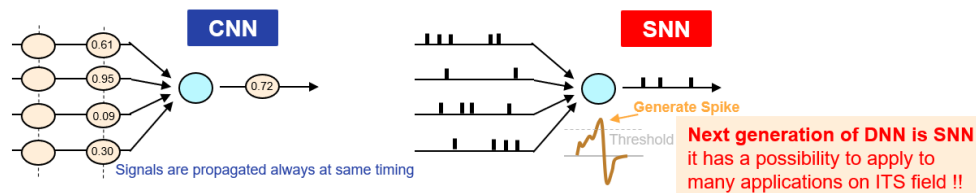


Figure 65 Difference between CNN and SNN

Combination between SNN and event sensor

Addition to SNN, we are also focusing on a sensor named event sensor or dynamic vision sensor (DVS) which was inspired by human retina ability. This sensor generates spike signals only when it observes a change in the intensity of the pixel. Therefore, in the scene that does not have much events, we can say this sensor is more energy efficient than RGB sensor because redundant data, which has no change between frames, is removed. Moreover, data format of the event sensor is suited for the input of SNN. Therefore, we think that the combination between SNN and event sensor is one of the best solutions to develop energy efficient system. In addition to the redundancy removal, the event sensor has other advantages such as high temporal resolution which could achieve 1 Mframe/s and high

dynamic range (> 120 dB). Thus, we feel a potential that event sensor might be a next generation sensor for ITS field.

How to train SNN or convert ANN to SNN

On the other hand, SNN is facing a difficulty in network training due to the non-differentiable characteristic, that is to say, conventional back-propagation cannot be applied straightforwardly to SNN training. In order to obtain well-trained SNN, many SNN training algorithms have been proposed. They are categorized into several approaches; the first approach employs modified back-propagation specific to SNN **Error! Reference source not found.** The second approach is spike-timing-dependent-plasticity (STDP), which is biologically inspired and is unsupervised learning algorithm **Error! Reference source not found.** Even though these two approaches reach same accuracy level as CNN in certain situation such as benchmark test on public database, there are not so many reports that applied them to general purpose. Addition to that, several points to be improved still exist, for example, training process takes long time.

Aside from the above approaches, there is one reasonable solution for SNN training, which converts trained CNN to SNN (CNN2SNN or ANN2SNN). In this method, at first, we train CNN with back-propagation, then, SNN is constructed by using weights of trained CNN and by modifying activation functions. In fact, we took this approach and developed a gesture recognition system using SNN on event sensor. As a result, we confirmed the potential of CNN2SNN method in terms of energy efficiency.

Issues of ANN2SNN

Through the development of CNN2SNN for gesture recognition, we found an issue that is performance decrease in accuracy after converting to SNN especially when a deeper architecture was employed. It is normal because error is accumulated when signal propagation increases.

For making this technology practical, we need to clarify how much performance shift exist when we use deeper architecture. Therefore, in this research, we measured the performance shift with changing depth of architecture.

4.4.3. Scientific contributions

As explained in previous section, we investigated performance shift between CNN and SNN converted from CNN. Experiments have been done on our hand gesture recognition dataset acquired by event sensor. We computed CNN training with changing architectures, then measured accuracies of converted SNN.

Through this investigation, we found the following points:

- The deeper an architecture is, the larger amount of accuracy decreases after SNN

conversion

- Optimal depth for converted SNN is different from CNN's one

The rest of this section is organized as follows: Next sub-section is dedicated to related works especially in CNN2SNN method. Then, our previous work about hand gesture recognition system using SNN on event camera is introduced. After that, experimental results of the performance shift after SNN conversion are described. In the final section, we summarize the results and conclude about the future of SNN based computer vision applications.

Related works of ANN2SNN

As of today, several researches describing CNN2SNN method have reported their performances and effectiveness in terms of computation cost compared with CNN [56][62]. In this technology, basic strategy is, at first, CNN is trained with back-propagation, then, SNN is constructed by sharing weights with trained CNN. To understand more details, we should explain the following points.

Spike encoding

In order to feed spike signals, we need to encode input signal as specific spike data format, which is depending on a SNN implementation. This encoding leads to an easy conversion from CNN to SNN. Several spike coding approaches have already existed such as rate coding, time coding, rank coding [64][65]. For example, in the rate coding, spike signals are generated by selecting the interval of spikes dt randomly from an exponential distribution with the parameter λ . The λ is computed by the following equation.

$$\lambda(x) = \frac{1}{\lambda_{max}^{-1} + (\lambda_{min}^{-1} - \lambda_{max}^{-1}) * (1 - x)} \quad (11)$$

In the equation, λ_{max} and λ_{min} are determined by normalized input intensity $x \in [0,1]$.

Integrate and fire neuron model

One of the biggest differences from CNN is that SNN employs Integrate and Fire (IF) neuron instead of conventional neuron model that does not have a function to accumulate input signals [66]. In the IF neuron, membrane potential U is computed to determine whether to emit a spike or not. Under the spike simulation time $t \in [0, T]$, membrane potential U_i^l is computed as below.

$$U_i^l(t + \Delta t) = U_i^l(t) - \theta s_i^l(t) + \sum_j w_{ij}^l s_j^{l-1}(t) \quad (12)$$

where w_{ij}^l is the weight between neuron j of layer $l - 1$ and neuron i of layer l . θ is a threshold to determine to generate a spike or not. The activation function $s_i^l \in \{0,1\}$ is a trigger function depending on the membrane potential U_i^l . A spike is generated if $U_i^l > \theta$, and then, U_i^l is reduced by an amount of θ .

Conversion learning

Before applying the IF neuron, propagated float value from activation function such as ReLU in CNN has to be converted to firing rates such as spike frequency in SNN. This conversion is done easily by scaling the weight matrices[67]. In the last layer whose neurons are associated to each class, prediction result is determined by the number of receiving spikes.

Hand gesture recognition system

As mentioned in the previous section, we developed a hand gesture recognition system by combining SNN with event camera for an in-cabin application. Under in-cabin situation, normally there is not so many events compared to outside of car, especially when car is moving. As another point, high dynamic range for operating day and night without lighting is required. Then, it goes without saying, low energy consumption is mandatory to install on vehicles. Considering these requirements, we decided to utilize the benefits of combination between SNN and event sensor. By realizing this kind of application that enables long-term operation with low energy consumption, we are planning to deploy it to intelligent vehicle and MaaS in future.

Training of CNN and inference by converted SNN

After generating the count images, CNN is trained as shown in the training part of Figure 66. Taking into account trade-off between gesture recognition accuracy versus real-time computation, we employed LeNet architecture for our hand gesture recognition system. The training was done using the N2D2 library [62]. Details for training are, number of iteration is 6 million times with a batch size 128, learning rate starts at 0.05, with decay= 5×10^{-4} , and momentum=0.9 with Adam optimizer momentum=0.9, and activation function is ReLU with a weight initialization of HeFiller approach [63].

For inference, at first, count-images are created as same as training process, then, they are converted to spikes which are propagated in SNN, as shown in the inference part of Figure 66. As you can see in the Figure, we generate synthetic (fake) spikes to recreate the stream of spikes from the count-image. In the recreation process, we use the rate coding to encode the value of the input pixel as the frequency of the spikes.

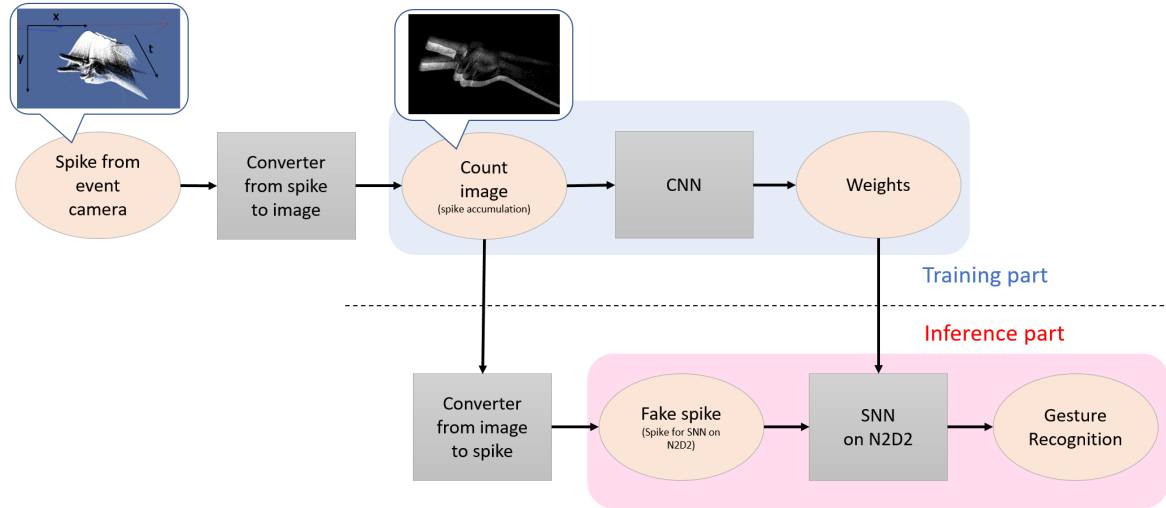


Figure 66 Processing flow for training and inference in hand gesture recognition system

Gesture recognition results and issues in SNN conversion

The result of gesture recognition is shown in Table 8. In this table, “energy gain” represents that computation cost of SNN is about 25 times less than that of CNN. Moreover, when we tuned the parameter λ_{min} for decreasing density rate of fake spike, the energy gain reached about 40 $\times$.

In contrast, recognition accuracy decreased from 93.6% to 89.4% after SNN conversion. This performance shift definitely came from conversion process, especially, the number of layers is one of the points that decreases the performance. In fact, LeNet is composed of two convolutional layers, two pooling layers and two fully connected layers. However, when we tested a deeper architecture than LeNet, its performance decrease was larger compared with the case of LeNet in the past experiments.

Table 8 Comparison between CNN and SNN

Model	Acc (%)	Energy gain ($\times$)
LeNet (CNN)	93.6	1
LeNet (SNN)	89.4	24.8

Investigation of performance shift after SNN conversion

In order to identify limitations of CNN2SNN method, we conducted the investigation of performance shift by SNN conversion. This investigation is important because we can judge whether CNN2SNN method can be applied or not by seeing an architecture. In the case that a system to be developed requires low energy consumption, we hope that this investigation results will be one of the references for designing a neural architecture to be converted to SNN.

The way of investigation is, to measure the performance shift with changing the depth of architecture. Concretely, we trained seven kinds of CNN which include from two to six convolutional layers with feeding same training dataset. Finally, we can see the relationship between performance shift and number of convolutional layers.

Prepared CNN architectures

Figure 67 shows CNN architectures that were used in this experiment. As you can see in the Figure, all architectures are composed of convolutional layer, max pooling layer and fully connected layer.

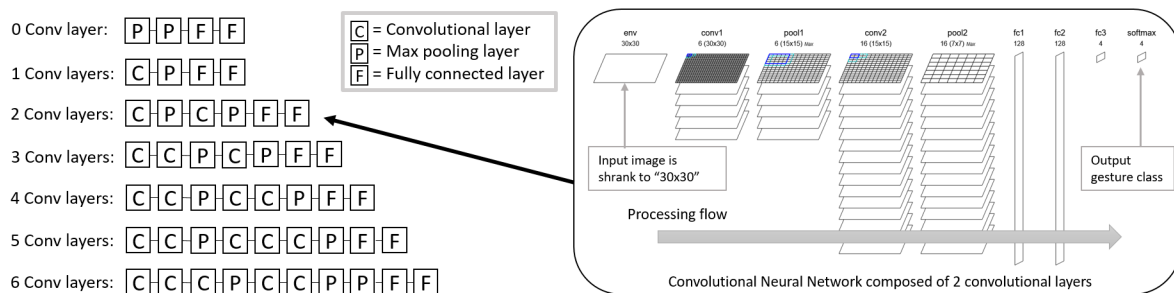


Figure 67 Prepared CNN architectures for performance shift investigation.

Dataset for hand gesture recognition

What we did concretely in the hand gesture recognition system development is, at first, to make a dataset recorded by event sensor. We gathered 3-class hand gesture (paper, rock and scissor) and background data, therefore totally 4-class dataset was constructed for training, validation and testing data. As we employed CNN2SNN method, frame-based images have to be prepared for CNN training. For this purpose, captured event streams (left-hand side and middle of Figure 68) are converted to "count-images (right-hand side of Figure 68)" whose pixel value represents the number of spikes arrived during a given time period (~few msec).

In this experiment, we constructed the count-image by accumulating events during 10 ms. After creating the count-image, we conducted data augmentation by shifting and rotating the count images. Finally, dataset is composed of 60.000 images in the training dataset, 6.000 images in the validation dataset, and 12.000 images in the test dataset.

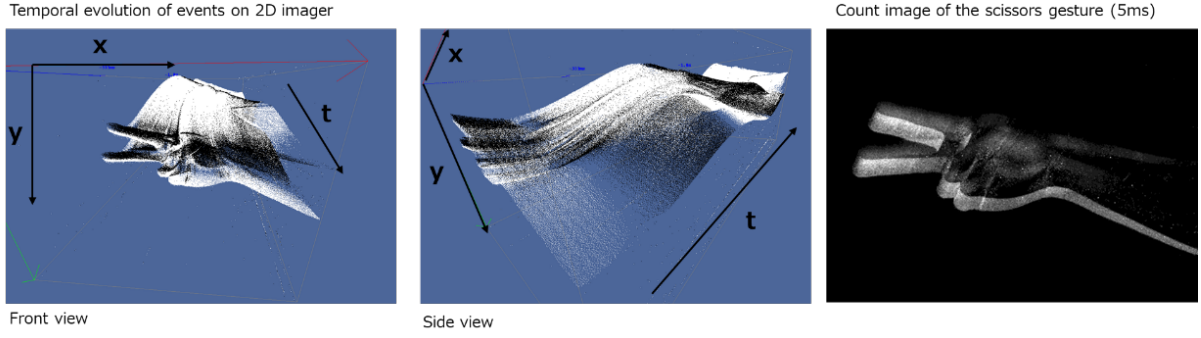


Figure 68 (left) Temporal evolution of events. Positive events (new changes) are in white and negative (disappearing) ones are in black. (right) Count-image of accumulated events within 5ms.

Experimental results

As described above, we trained seven kinds of CNN which are shown in Figure 67 with feeding the hand gesture dataset captured by event sensor. In each architecture, training was conducted 10 times with changing initial weight values, therefore, we had totally 70 results in this experiment. All the experiments have been conducted by a laptop with the following configuration: core i7, 16GB RAM and graphic board 1070 GTX.

Figure 69 includes a boxplot of performance shift after SNN conversion and a chart for recognition accuracy of trained CNN and converted SNN as a function of the number of convolutional layers. In the boxplot, the performance shift (PS) is calculated as follow,

$$PS = Acc_{CNN} - Acc_{SNN} \quad (13)$$

where Acc_{CNN} and Acc_{SNN} are accuracy of CNN and SNN on test data, respectively. Horizontal axes of both the boxplot and the chart are number of convolutional layers.

At first, you can see in the boxplot that the performance shift was larger when the number of convolutional layers increased. From this result, we can say that we have to take care when we apply CNN2SNN method to deeper architecture. At the same time, the figure shows that the best (lowest) performance shift is the architecture composed of two convolutional layers.

In our analysis, one of the reasons for this result is included in the trend of accuracy of CNN and SNN that is shown in right-hand side of Figure 69. In CNN, the accuracy increases significantly from zero to two convolutional layers, then it increases moderately. On the other hand, accuracy of SNN has a same trend as CNN until two convolutional layers, however, it turns to decrease if convolutional layers are added. Though CNN2SNN aims to minimize errors in conversion of neuron model and so on, perfect conversion method that has no error in conversion does not exist for the moment. In fact, the error in conversion accumulates, as a result, performance shift increases. Based on this point, we can say that the risk for the errors increases if architecture becomes deeper. This is because deeper network requires a lot of conversion process. This is the reason for accuracy decrease in deeper SNN. In contrast, we know deeper CNN has better performance in general. It means trade-off relationship in depth

of architecture for CNN2SNN exists. Therefore, we suggest that to start trial for architecture determination from shallow one, then to stop adding layers if the accuracy of SNN turns to downward trend even if the accuracy of CNN continues to increase.

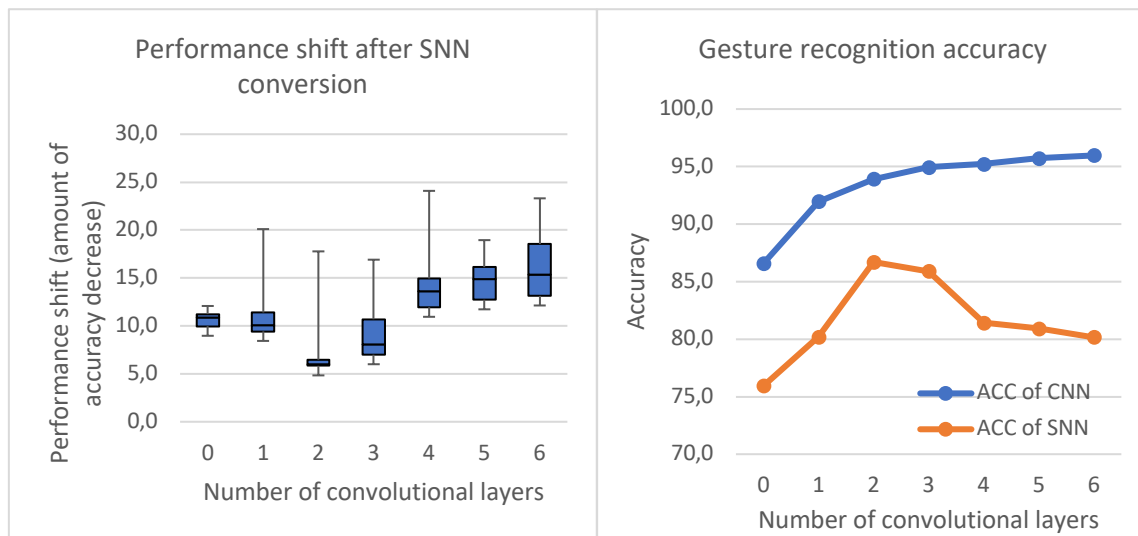


Figure 69 Boxplot for performance shift after SNN conversion and chart describing gesture recognition accuracy of CNN and SNN

It might be more interesting if you see further analysis about the performance shift depending on the number of convolutional layers in Figure 70. This figure describes a chart of standard deviation (SD) from iterative experiments of performance shift (10 samples in each architecture). As shown in Figure 69, we can see that architecture including two convolutional layers has lowest performance shift. In Figure 70, we are able to draw the same conclusion. Furthermore, the line of the standard deviation shows clear trend indicating that two convolutional layers is best. It means that the standard deviation could become an indicator to determine an architecture in CNN2SNN method because low standard deviation leads to more stable conversion, as well as the timing of conversion in CNN trading described above.

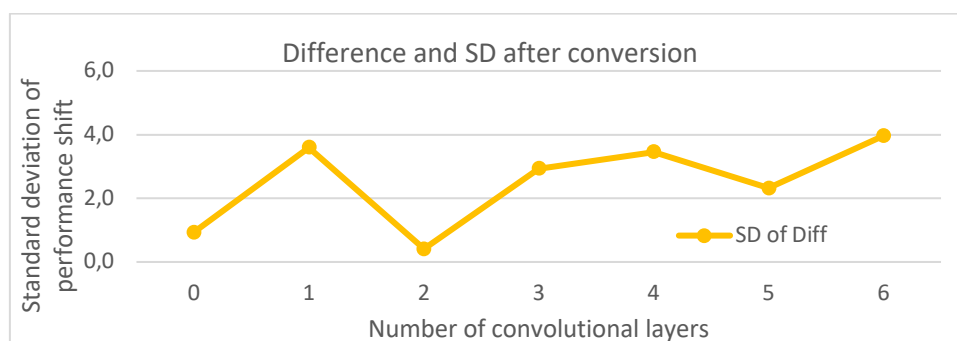


Figure 70 Standard deviation of gesture recognition accuracy in each architecture

Conclusion

In this work, we investigated a performance shift of the method to convert CNN2SNN. From the experimental results, we confirmed, basically, the deeper neural architecture is, the larger the performance shift after SNN conversion becomes. Another interesting point, is that optimal number of layers to reach maximum classification is different between CNN and SNN. As our suggestion, the trends of accuracies in CNN and SNN as a function of number of layers, and standard deviation of performance shift in iterative experiments might be indicators for suitable neural architecture design.

In order to make CNN2SNN a reliable method, we will continue further analysis. In future works, we will analyze performance shift derived from spike generation method such as spike encoding algorithm and its parameters. We will apply our method to other datasets to clarify the behavior of the method with more complex tasks. We will also evaluate more complex architecture such as ResNet (specially to establish a method to convert to SNN the residual part of ResNet). We will not only clarify how to choose the suitable SNN depth but also its width. This will be done by designing and training wider CNNs and convert them.

Through the research activity for SNN, we are going to make a computer vision-based application with high classification performance but with very low energy consumption. Our aim is allowing this technology to be a game changer for next generation of applications for CASE and smart city. We believe that products using SNN will lead new era of automotive and many other domains.

4.5. Summary

This chapter has presented my first research and development experience within the automotive domain, which represented a big transition from my prior work in the medical field. At the outset, the adaptation process was far from trivial. The diversity of sensor modalities and the wide range of targeted applications introduced a kind of new world to me. Nevertheless, once familiarity was established, the novelty of these challenges transformed into an intellectually stimulating and rewarding endeavor.

A particularly significant difference from the medical domain was the abundance of data. Whereas medical applications are constrained by stringent privacy regulations that severely limit data acquisition, the automotive domain provided access to vast and various datasets. This abundance not only facilitated the training of artificial intelligence models but also allowed for exploration of state-of-the-art neural network architectures. In turn, this accelerated experimentation, validation, and refinement of advanced AI methodologies.

Another crucial aspect was the expansion beyond image-centric data. Engaging with non-visual modalities—such as time-series sensor or multimodal data—demanded the development of dedicated representational structures and domain-specific processing

techniques. This requirement for specialized solutions promoted a deeper appreciation for the interplay between algorithmic design and application context. Moreover, it unveiled new opportunities for innovation, both at the academic and industrial levels, by suggesting pathways to novel architectures and business models tailored to unique data characteristics.

In summary, the automotive R&D experience not only broadened my technical expertise across diverse sensor types and application scopes but also reshaped my perspective on the strategic role of data availability, modality-specific architectures, and the discovery of new opportunities at the intersection of AI research and industry deployment. This chapter thus highlights the transformative nature of this domain shift, underscoring its long-term impact on my research trajectory.

Chapter 5

Future plan

5.1. Introduction

The experimental results presented in Chapter 4 showed the effectiveness of SNNs in terms of low power consumption and asynchronous processing. These findings extend beyond academic significance and clearly demonstrate practical potential of industrial research and development. Consequently, within our research organization, SNN technology has been positioned as a core competence for next-generation R&D activities. In particular, in my role as research manager, I recognize SNNs as a strategic pillar to be advanced over the coming years.

This chapter discusses two major research directions related to the application of SNN technology. The first concerns low-power object detection, a line of study that has already been initiated. This research aims to energy enable efficient object recognition in edge devices and mobile environments, where the energy efficiency of SNNs can be fully leveraged. The second direction, which is planned for future development, involves multimodal signal processing exploiting the asynchrony of SNNs. This approach seeks to establish a new framework for integrating and processing various kinds of signals—such as image, wave (sound, current, etc), point cloud data—in an efficient and biologically inspired manner. Such a direction is expected to hold significant relevance for next-generation applications.

It should be noted that both projects are being conducted within the corporate research environment, and thus, disclosed contents in this chapter is limited by confidentiality constraints. Therefore, this chapter explained mainly our future directions and show some concrete technologies and results that have been already disclosed in public.

Although from Chapter 2 to 4 show the achievements that were described mainly in the papers I (Yuta Nakano) authored as the first author, the contents in this chapter were disclosed by the paper of our team. Thanks to the team members, we are accumulating good results day by day.

5.2. SNN object detection with ultra-low energy consumption

As a first target application in the project using SNN, we started to develop super low energy consumption system for object detection, especially pedestrian detection because various applications that our headquarter is planning to develop requests it. As we all know, pedestrian and other object detection on RGB data by conventional Deep Neural Networks have already achieved substantially high level. However, their energy consumptions prevent sometimes installing them as edge devices. This is very important point for us because we are

developing systems for vehicle, mobility, robot which are often facing battery problem: how to reduce energy consumption even if we add new functions. This is our motivation to struggle with low energy consumption system for object detection by using SNN.

5.2.1. Background and conventional methods

Object detection is a fundamental task in computer vision, and pedestrian or human detection plays a critical role in applications such as autonomous driving, surveillance, and robotics. While conventional deep learning approaches have achieved remarkable accuracy, they often require substantial computational resources and power consumption, which poses limitations for edge devices and real-time deployment. To overcome these challenges, recent studies have actively explored the use of SNN for object detection, leveraging their event-driven and energy-efficient characteristics.

One major line of research focuses on event-camera-based object detection. Li et al. [69] proposed Spiking-YOLO, which processes event streams for low-latency and energy-efficient pedestrian detection. Their results demonstrated that SNNs provide more stable performance than frame-based CNNs in high-speed scenarios.

In the area of real-time pedestrian detection, Zhou et al. [70] developed an SNN-based detector compatible with Intel Loihi, demonstrating superior energy efficiency in vehicular environments. Similarly, Park et al. [71] introduced a hybrid SNN model combining event streams with spike-based encoding, achieving significantly higher power efficiency compared to conventional CNNs for pedestrian detection.

Furthermore, research has investigated multimodal fusion for human detection. Safa et al. [72] integrated event-camera and radar data using spike-timing-dependent plasticity (STDP) for continual learning, enabling robust detection and tracking of humans even under adverse conditions such as low illumination or inclement weather.

These studies collectively demonstrate that SNNs hold strong potential for object detection, particularly for pedestrian recognition, by exploiting low-power and event-driven processing. Nevertheless, challenges remain in balancing detection accuracy with inference speed and addressing hardware constraints in real-world deployment. The present study aims to tackle these challenges and to establish an SNN-based framework for next-generation pedestrian detection on edge intelligence platforms.

5.2.2. Our current achievements

Up to now, we have already published several papers about super low energy consumption system for object detection. All of them, we use SNN and neuromorphic hardware. The following sub sections shows our current achievements

Real-time and low power pedestrian detection

Many researchers in this domain describe on their papers that event cameras provide asynchronous, sparse, and energy-efficient data acquisition, while Spiking Neural Networks (SNNs) promise low-power, biologically inspired processing. Therefore, these technologies are a kind of best-combination for realizing super low energy vision system. However, addition to these technologies, neuromorphic hardware who executes SNN is mandatory to make a real product. Even though we know this fact, most prior SNN object detection works remain at simulation stage, with very few deployed on neuromorphic hardware. Our team, with H.Bulzomi at its core, addresses the gap by proposing a fully asynchronous, end-to-end SNN system for pedestrian detection implemented on neuromorphic hardware[68]. To be more specific, we showed the following technical contributions.

Technical contributions

1. First anchor-free real-time pedestrian detection system using an SNN.
2. Introduction of a novel weighted loss function that scales importance of objects based on the number of events inside their bounding boxes, improving detection robustness.
3. Integration of a regularization term in the loss to minimize synaptic operations (SynOps), thus reducing power consumption and latency.
4. Demonstration of a fully on-chip implementation, validating the feasibility of real-world deployment in energy-constrained edge environments.

Although you can see details in the paper published in CVPR workshop[68], I just explain several above points.

Synaptic Operation Minimization

The efficiency of SNNs largely stems from their sparse activations, which significantly reduce computational costs compared to conventional ANNs. To further exploit this property, our team introduces a sparsity regularization term into the loss function, explicitly encouraging the minimization of synaptic operations (SynOps) during training. The regularization is formulated as:

$$L_{SynOps} = \sum_{l=1}^L ReLU\left(\frac{SynOps_l - TargetSynOps}{TargetSynOps}\right) \quad (14)$$

where $SynOps_l$ denotes the number of synaptic operations at layer l , L the total number of layers, and $TargetSynOps$ an empirically chosen threshold representing the maximum sustainable operations for real-time inference. The ReLU ensures penalties are applied only when the SynOps exceed the target, avoiding unnecessary constraints.

This formulation directly links network activity to hardware efficiency, reducing redundant computation, accelerating on-chip inference, and lower energy consumption. Experimental results confirm that this regularization decreases SynOps by over 97%, while maintaining competitive detection accuracy, making it highly suitable for edge deployment.

Model architecture using Anchor-free method

The proposed model adopts an anchor-free detection paradigm, inspired by CenterNet and prior work by Liu et al[69]. Unlike anchor-based detectors such as YOLO or SSD, which generate multiple candidates bounding boxes per location followed by a costly non-maximal suppression (NMS) step, the anchor-free approach directly predicts object centers and scales. This design is particularly well-suited to neuromorphic hardware, as it reduces the number of output neurons and avoids additional computations outside the chip.

The architecture illustrated in Figure 71 is a lightweight fully convolutional SNN backbone with detection heads dedicated to center and scale prediction. Batch normalization layers are used during training but are fused into convolutional layers before porting the model onto the Speck chip, ensuring compatibility with hardware constraints. Training is performed using the surrogate gradient method, where sparse asynchronous events are converted into dense frames for gradient-based optimization. After training, inference on-chip processes raw event streams asynchronously, preserving the energy efficiency of the SNN while maintaining accurate pedestrian detection.

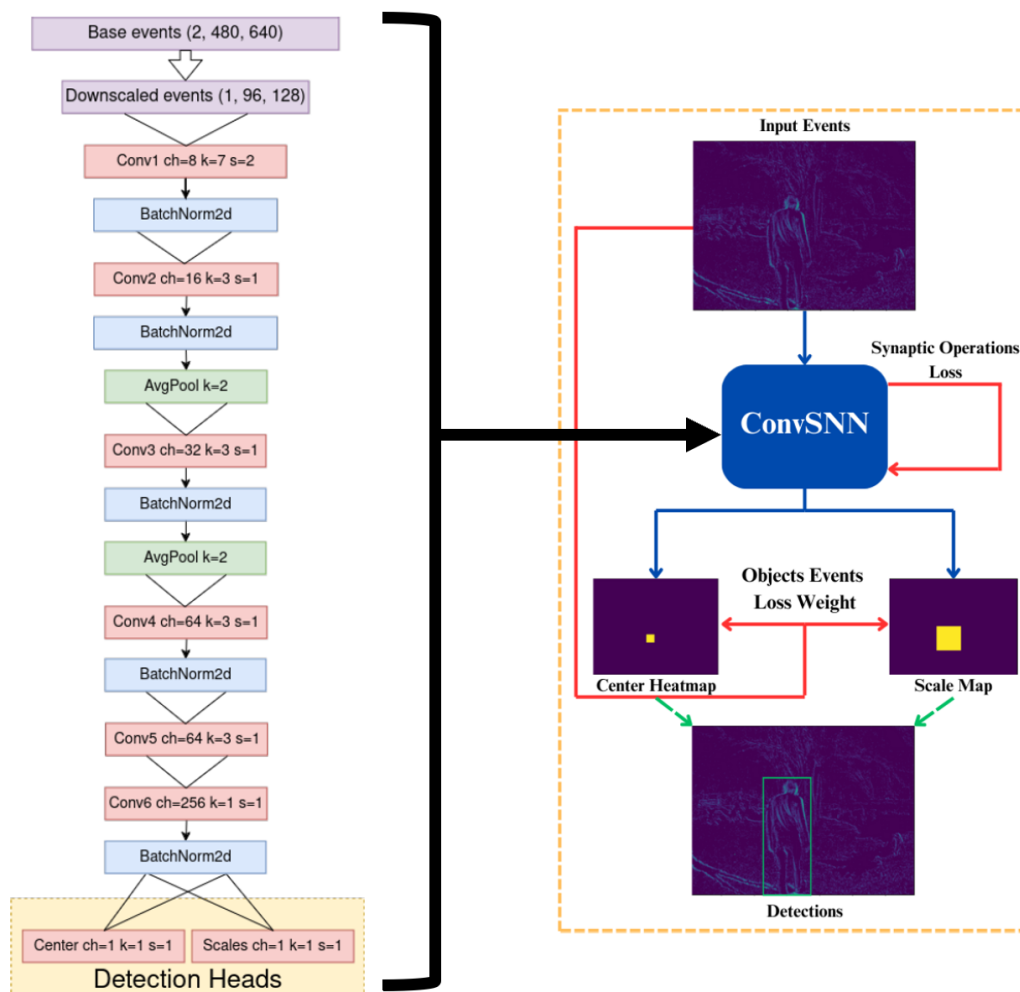


Figure 71 SNN architecture for pedestrian detection

Experimental results

The experimental evaluation focuses on assessing the impact of the proposed loss components on pedestrian detection performance. An ablation study of the event-based weighting term shown in Table 9 demonstrates clear improvements: when applied to the SNN version of the model, the F1 score increased by +2.2 points (from 0.687 to 0.709) and the mean Average Precision (mAP) improved by +1.7 points (from 0.319 to 0.336). Similar but smaller gains were also observed in the ANN counterpart, confirming that the weighting mechanism facilitates better feature learning across event-rich objects. Importantly, the method improved detection consistency across all object categories (few, medium, and many events).

Ablation of the SynOps regularization shows that it reduced synaptic operations by 97.3% (1161M $\rightarrow$ 29M SynOps/s) and decreased inference latency from 216 ms to 24 ms, while only modestly reducing accuracy (F1: -3.67%, mAP: -10.4%).

Table 9 Ablation study of our weight term in both the SNN and equivalent ANN version of our network

Results with mAP[0.3] on PEDRo dataset				
Model	f1	Precision	Recall	mAP
SNN without weight term	0.687	0.67	0.704	0.319
SNN with weight term	0.709	0.722	0.709	0.336
ANN with weight term	0.746	0.736	0.756	0.352
ANN without weight term	0.749	0.74	0.759	0.358

Overall, the results highlight a favorable trade-off: the proposed network maintains competitive accuracy while drastically reducing computation and energy requirements, validating its suitability for real-time edge deployment on neuromorphic hardware.

Lane detection

Aside from pedestrian detection, we are also struggling with other type of processing such as semantic segmentation on event data with SNN, concretely, lane detection on event sensor installed on vehicle. The lane detection is a critical component of autonomous driving systems, often formulated as a semantic segmentation task. Conventional CNN-based approaches (e.g., U-Net, FCN) achieve strong performance but are computationally heavy and unsuitable for real-time inference on energy-constrained embedded automotive platforms. Event cameras combined with SNN and neuromorphic hardware provide a promising path toward low-power lane detection. Previous work (e.g., LaneSNN on Loihi [73][74]) demonstrated feasibility, but models remained relatively complex and energy-hungry. This study introduces a compact, ultra-low-power SNN architecture for lane detection. The

achievements in this study were published by C.E.Mauriello who is a core member of this project[75].

Technical contributions

In this study, we showed the following technical contributions:

- Efficient Preprocessing: Cropping and resizing strategies adapted to hardware limits (final input: 128×128 pixels), plus probabilistic and copying frame construction methods for event data.
- Design of a lightweight SNN (about 6K parameters) for semantic segmentation of lanes, drastically smaller than previous models
- Deploy lane detection on neuromorphic hardware with real-time capability, and with 12 mW power consumption which is two orders of magnitude lower than prior art.

About the preprocessing, its details are shown in Figure 72. Then, for the architecture of SNN, our SNN follows an encoder–decoder structure with five convolutional layers in the encoder, gradually increasing from 4 to 32 channels, and spiking Integrate-and-Fire (IAF) activations to capture temporal dynamics. A lightweight convolutional decoder outputs 32×32 lane segmentation maps aligned with pre-processed labels, resulting in a network with only 6365 parameters. To stabilize predictions, outputs are averaged across 30 temporal frames, providing robust lane detection while remaining compatible with neuromorphic hardware constraints.

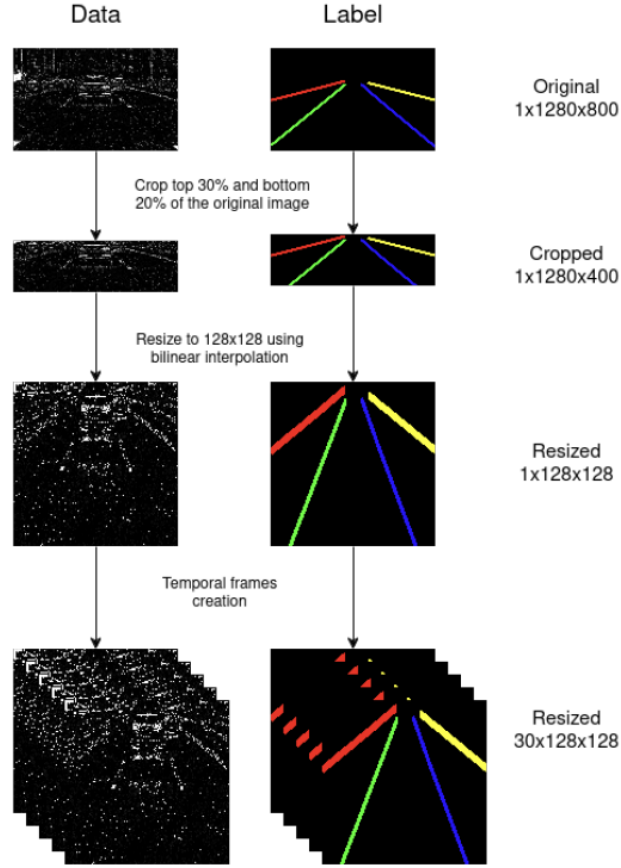


Figure 72 Preprocessing in our lane detection

Experimental results

The proposed SNN-based lane detection models were evaluated both in software and on neuromorphic hardware using the DET dataset whose examples are shown in Figure 72. Performance was assessed with the Intersection over Union (IoU) metric, along with latency and power consumption.

Results showed that fully convolutional network (FCNN) variants were more suitable for neuromorphic deployment compared to fully connected decoders, which achieved slightly higher IoU scores (up to 0.3103) but exceeded memory limits and were therefore not deployable. Among the deployable networks, the bias-free FCNN (FCNN NB) achieved an IoU of 0.20 on hardware with an inference latency of 19.26 ms over 30 frames.

Table III compares this approach against state-of-the-art LaneSNN implementations. While existing models such as CNNLaneSNN and C600 LaneSNN reached IoU scores of 0.6 with over 1M parameters and consumed approximately 1 W, the proposed FCNN NB required only 6k parameters and 12 mW of power, achieving IoU scores of 0.3 in software and 0.2 on hardware.

These results underline a fundamental trade-off between accuracy and efficiency, demonstrating that the proposed architecture drastically reduces energy consumption while

maintaining sufficient performance for real-time lane detection in embedded automotive systems.

5.2.3. Future plan

In above sections, we demonstrated the feasibility of building a low-power, real-time object detection system by combining event cameras, Spiking Neural Networks, and neuromorphic hardware. This approach highlights the potential of bio-inspired computing for embedded perception tasks where energy efficiency and low latency are critical. Looking ahead, our next steps will focus on broadening the scope of detection targets and exploring new algorithms that can further improve accuracy while maintaining hardware compatibility.

Another promising direction is the extension of SNN-based methods to a wider variety of sensory inputs. While this study concentrated on event-based vision, we are aiming in future to adapted this technology to other modalities such as traditional vision sensors, milli-wave radar, or LiDAR, each of which provides complementary information in different formats. Developing architectures and training strategies capable of fusing such heterogeneous inputs would significantly enhance robustness and versatility. Ultimately, these efforts aim to establish neuromorphic systems as a general, multimodal framework for efficient real-time perception in safety-critical applications for automotive and robotic domains.

5.3. Multi-modal SNN: handling various sensor data simultaneously

As I explained in previous section, we focused at first the sparsity and energy efficiency of SNN, and conducted several preliminary studies and technologies developments. Through these activities, we also recognized the importance and potential of asynchrony of SNN, which can deal with various sampling rate and time series data. In our automotive and industrial context, we handle not only vision data but also voltage and current waveform data for motor control, various data from heterogeneous equipment in factory, etc. Addition to this variety, we have to sometimes process multiple kinds of signal at the same time, for instance, we have to take into account kilo Hz order data and asynchronous event simultaneously. I think that the asynchrony of SNN can give a solution for them and I'm now making plans of this idea "multi-modal SNN". In this section, I will show them briefly.

5.3.1. Prior arts of SNN for various modalities

Here, we show representative prior studies that exploit the asynchrony of SNN for data modalities beyond vision.

In the domain of audio–visual (AV) integration, substantial progress has been achieved by exploiting the temporal coding and asynchrony of SNNs. Liu et al. proposed MISNet [76], which enables cooperative integration of audio and visual modalities with high accuracy and low energy consumption. Li et al. developed a Spiking Tucker Fusion Transformer for zero-shot learning [77], while He et al. enhanced AV SNNs through semantic alignment and cross-modal residual learning [78]. Furthermore, Guo et al. reported an AV classification framework that integrates cross-modal attention with SNN–Transformer architectures [79]. Collectively, these studies demonstrate the effectiveness of SNNs in extending conventional AV fusion frameworks.

In the area of visual–tactile integration, research has been particularly active in robotics applications. Taunyazov et al. combined event-driven vision sensors with tactile sensing, enabling efficient multimodal processing on neuromorphic hardware such as Intel’s Loihi [80]. More recently, Yang et al. introduced temporal–location spiking neurons that capture both temporal and spatial features, thereby improving the performance of visual–tactile fusion [81].

Research in radar–event camera integration has also demonstrated the utility of SNNs, particularly leveraging spike-timing-dependent plasticity (STDP) learning for sequential and energy-efficient computation. Safa et al. fused information from event-based cameras and FMCW radar for SLAM applications [82], while Vogginger et al. showed the applicability of SNNs in automotive radar processing pipelines [83].

In addition, several works have addressed extensions of learning algorithms for multimodal SNNs. Rathi and Roy proposed an STDP-based unsupervised cross-modal learning framework for simultaneous processing of images and audio [84]. More recent approaches include Li et al.’s SpikeCLIP, which distills multimodal knowledge into SNNs [85], and He et al.’s temporal-attention-guided SNNs that integrate spiking computation with Transformer-based architectures [86].

Taken together, prior work highlights the potential of SNNs to overcome the limitations of conventional multimodal learning approaches. Nevertheless, several challenges remain unresolved, including integration strategies, stability of learning, and the balance between energy efficiency and accuracy in practical scenarios. The present study aims to address these challenges and to establish a foundation for the practical deployment of multimodal processing frameworks based on SNNs.

5.3.2. Future plan

Over the next several years, I will pursue a sustained program of research and development on multi-modal spiking neural networks (SNNs). Working from within the automotive domain, it means to deal with vehicles, mobility services, and factory robotics, I see unique promise in SNN architectures that can ingest and reason over heterogeneous sensor streams concurrently and efficiently. My long-term ideal, as envisioned in Figure 73, is a brain-

inspired system in which a single SNN processes diverse signals end-to-end, enabling robust perception and control across real-world operating conditions.

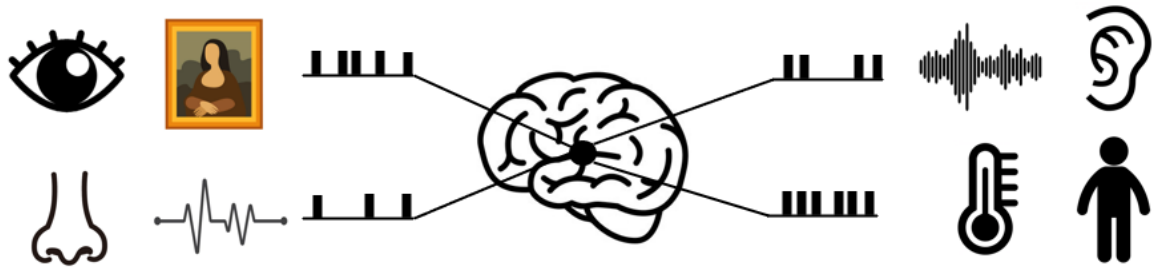


Figure 73 Idea of multi-modal SNN

I recognize that this goal is ambitious. Even at the level of signal transduction, the spike-encoding front end lacks a widely accepted standard; converting disparate sensor outputs into informative spike trains, while preserving timing, scale, and cross-modal alignment, remains an open problem. Yet I am committed to addressing these challenges deliberately, by iterating on encoding schemes, advancing learning rules that exploit temporal structure, and validating designs on representative automotive and robotics tasks. Step by step, I aim to transform these difficulties into opportunities for innovation, ultimately delivering technologies that raise safety, efficiency, and adaptability in mobility and industrial environments. By solving these problems one by one, I hope to contribute not only to our field, but also to other industries and finally to society at large.

Chapter 6

Curriculum Vitae

General details

First name: Yuta

Last name: NAKANO

Status: Head of Computer Science department in IMRA EUROPE /

Co-director of Ph.D thesis in CIFRE framework

Institution: IMRA EUROPE

Birth date and place: 25/05/1980 in Gumma (Japan)

Post address: 220 Rue Albert Caquot, 06560 Valbonne

Email address: nakano@imra-europe.com

Phone: +33 7 81 17 90 96

Career

- Since Oct. 2018: IMRA EUROPE
 - Since Apr. 2025: Head of Computer Science dpt. at IMRA EUROPE
 - Jan. 2024 – Mar. 2025: Deputy Group Manager of Computer Science dpt.
 - Apr. 2021 – Dec. 2023: Project leader
 - Oct. 2018 – Mar. 2021: R&D researcher
- Since Nov. 2023: Co-director of PhD students
- Jun. 2016 – Sep. 2018: Lead scientist at MEDIAN Technologies
- Apr. 2008 – May. 2016: R&D engineer at CANON Inc.
 - Feb. 2012 – Feb.2015: Researcher at Canon Research Centre France
- Apr. 2005 – Mar. 2008: PhD in Computer Science at Yokohama National University, Japan
 - Thèse de doctorat « A study of Evolutionary Medical Image Processing»

Research directions

Teaching activities / PhD students supervision

- Mr. Hugo Bulzomi (Université Côte d'Azur) (50% with Prof. Jean Martinet), since Nov. 2023
 - Spiking Neural Network for object detection on event sensor data
 - Delay learning on Spiking Neural Network
- Ms. Sadia Memudu (Université Côte d'Azur) (50% with Prof. Jean Martinet), since Jan. 2025
 - Object detection on ego-motion sensor data by using Spiking Neural Network

Publications

- Conférences internationales (papiers et présentations orales)
 - Crescenzo Edoardo Mauriello, Hugo Bulzomi, Jean Martinet and Yuta Nakano : End-to-End Neuromorphic Lane Detection; International Workshop On Advances in Sensors and Interfaces (**2025**)
 - Hugo Bulzomi, Sadia Memudu, Yuta Nakano and Jean Martinet : Real-Time Pedestrian Detection at the Edge on a Fully Asynchronous Neuromorphic System; Conference on Computer Vision and Pattern Recognition workshop on Event-based Vision (**2025**)
 - Hugo Bulzomi, Yuta Nakano, Rémy Bendahan and Jean Martinet: Neuromorphic Temporal Pattern Detection with Spiking Neural Networks using Synaptic Delays; International Conference on AI Circuits and Systems (**2024**)
 - Hugo Bulzomi, Yuta Nakano, Rémy Bendahan and Jean Martinet : Object Detection for Embedded Systems Using Tiny Spiking Neural Networks: Filtering Noise Through Visual Attention; International Conference on Machine Vision Applications (**2023**)
 - Yuta Nakano, Kozo Kajiwar, Atsushi Hori and Takeshi Fujita : Event Data Representation based on Time Stamp for Pedestrian Detection; International Conference on Intelligent Vehicles and Applications, Prague, Czechia 03-04 July (**2023**)
 - Yuta Nakano, Takeshi Fujita et al. : Data representation of event data for object detection; Symposium on Sensing vis Image Information, Yokohama, Japan, 14-16 June (**2023**)
 - Frederic Abad, Yuta Nakano, Remy Bendahan, Akiyoshi Takabe et al. : Locally Masked Convolutional Models for Anomalous Sound Detection by Deep Learning; International Conference on Industry 4.0 and Smart Manufacturing, Lisbon, 22-24 November (**2023**)
 - Sylvain Bougnoux, Rémy Bendahan, Yuta Nakano et al. : Control of an Air-Cored Resonant Induction Motor ; International Conference on Electrical Machines, Valencia, Spain, 5-8 September (**2022**)
 - Yuta Nakano, Takeshi Fujita and Rémy Bendahan : Investigation of Performance Shift by Conversion from Convolutional Neural Networks to Spiking Neural Networks; ITS world congress, Hamburg, Germany, 11-15 October (**2021**)
 - Yuta Nakano: Non-invasive phenotyping of hepatic fibrosis using unsupervised machine learning; ESMO (**2018**)
 - Yuta Nakano, Estanislao Oubel, Johan Brag, Alice Doreille and Olivier Lucidarme: Quantitative imaging biomarker of disease progression in hepatocellular carcinoma and NASH; NASH workshop (**2017**)

- Yuta Nakano, Shinichi Shirakawa, Noriko Yata and Tomoharu Nagao: Automatic Construction of Image Transformation Algorithms Using Feature Based Genetic Image Network; Proc. of the 2010 IEEE Congress on Evolutionary Computation (**2010**)
- Yuta Nakano and Tomoharu Nagao : Automatic construction of moving object segmentation from video images using 3D-ACTIT; Proc. of The 2007 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2007), pp.1153-1158, October 7-10, (**2007**)
- Yuta Nakano and Tomoharu Nagao: Automatic construction of abnormal signal extraction processing from 3D diffusion weighted image; Proc. of IWAIT-2007, Bangkok, Thailand, Jan.8-9, P5-05 (**2007**)
- Yuta Nakano and Tomoharu Nagao: Automatic extraction of internal organs region from 3D PET image data using 3D-ACTIT; Proc. of IWAIT-2006, Okinawa, Japan, Jan.10 (**2006**)
- Yuta Nakano and Tomoharu Nagao: Automatic construction of internal organs segmentation from PET image data using 3D-ACTIT; Proc. of the Joint 2nd International Conference on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems, SCIS & ISIS 2004, September 21-24, 2004, Yokohama (**2004**)
- Yuta Nakano and Tomoharu Nagao: 3D medical image processing using 3D-ACTIT; Automatic Construction of Tree-structural Image Transformation; Proc. of the International Workshop on Advanced Image Technology IWAIT'04, January 12-13, 2004, Singapore (**2004**)
- Journal
 - Yuta Nakano and Tomoharu Nagao : Proposition of 3D-ACTIT and Its Application to PET image processing, Medical Imaging and Information Sciences, vol.24, No.4, pp.119-125 (**2007**)
 - Yuta Nakano and Tomoharu Nagao : Automatic construction of abnormal signal extraction processing from 3D diffusion weighted image, Journal of The Japanese Society of Medical Imaging Technology, Vol.25, No.5, pp.362-370 (**2007**)
 - Yuta Nakano and Tomoharu Nagao : Automatic extraction of liver regions from 3D-PET images using 3D-ACTIT, Journal of The Institute of Image Information and Television Engineering, Vol.62, No.3, pp.424-434 (**2008**)
- Brevets
 - Au moins 20 brevets ont été publicités par moi dans le domaine processus d'image et machine learning
 - e.g.
<https://docs.google.com/viewer?url=patentimages.storage.googleapis.com/pdfs/US8620048.pdf>
 - J'ai des expériences de investigation de 4,000 brevets pour lancer des produits
 - Dans ce projet, j'étais un chef qui a fait une stratégie d'investigation et

négocié avec département propriété intellectuelle

References

- [1] T. Anjuin, M. Nakajima, "Fundamentals of Image Engineering", Shoko-do, 1986
- [2] J. Toriwaki, "Digital Image Processing for Image Understanding [I][II].", Shoko-do, 1988
- [3] JI Pastore, EG Moler, VL Ballarin, "Segmentation of brain magnetic resonance images through morphological operators and geodesic distance", Digital Signal Processing, Vol.15, pp.153-160, 2005
- [4] S. Srivastava, F. Maes, D. Andermuellen, et al, "Feature-based statistical analysis of structural MR data for automatic detection of focal cortical dysplastic lesions", Neuro Image, Vol. 27, pp. 253-266, 2005
- [5] H. Kobata, "Vector Concentration Filter and Its Application to Medical Image Processing", Transactions of the Institute of Electronics, Information and Communication Engineers D-II, J87-D-II, No. 1, pp. 19-30, 2004
- [6] Y. Yoshinaga, H. Kobata, "A method for detecting line regions unaffected by contrast or width using brightness gradient vectors", IEICE Transactions on Communications, D-II, J81-D-II, No. 11, pp. 2547-2555, 1998
- [7] A. Shimizu, J. Hasegawa, J. Toriwaki, "Minimum Directional Difference Filter for Detecting Mass Shadows in Chest X-ray Images and Its Properties", IEICE Transactions on Communications, D-II, J76-D-II, No. 2, pp. 241-249, 1993
- [8] A. Shimizu, Y. Sato, "Construction of a Statistical Atlas of Abdominal Organs and Its Application to Multi-Organ Segmentation", Medical Imaging Technology, Vol. 24, No. 43, pp. 153-160, 2006
- [9] J. Hasegawa, H. Kubota, J. Torawaki, "Image Processing Expert System IMPRESS Using Sample Image Presentation Method", Information Processing Society of Japan, Vol. J70-D, No. 11, pp. 2147-2153, 1987
- [10] T. Matsuyama, M. Ozaki, "LLVE: An Image Processing Expert System for Top-Down Segmentation", Transactions of the Information Processing Society of Japan, Vol. 27, No. 2, pp. 191-204, 1986
- [11] H. Tamura, H. Sato, K. Sakakibara, Fumio Kubo, " DIA-Expert System and Its Knowledge Representation Method", Vol. 29, No. 2, pp. 199-208, 1988
- [12] J.H. Holland, "Adaptation in Natural and Artificial Systems", University of Michigan Press (1975)
- [13] J.R. Koza: "Genetic Programming: On the Programming of Computers by Means of Natural Selection", MIT Press, 1992
- [14] S. Iba, "Genetic Programming", Tokyo Denki University Press, 1996
- [15] S. Aoki, T. Nagao, "Automatic Construction of Tree-structural Image Transformation Using Genetic Programming", Journal of the Institute of Image Information and Television Engineers, Vol. 53, No. 6, pp. 888-894, 1999

- [16] S. Aoki, T. Nagao, "Automatic Construction of Tree-structural Image Transformation Using Genetic Programming", International Conference of Image Processing, Kobe, Japan, pp.529–533, 1999
- [17] S. Shirakawa, S. Ogino, and T. Nagao, "Automatic Construction of Image Transformation Using Genetic Image Network", Information Processing Society of Japan, Vol. 48, No. SIG 19, pp. 117-126, 2007
- [18] S. Shirakawa, T. Nagao, "Feed Forward Genetic Image Network: Toward Efficient Automatic Construction of Image Processing Algorithm", International Symposium on Visual Computing (ISCV 2007), Lake Tahoe, USA, pp.287-297, 2007
- [19] J. Toriwaki, "3-Dimensional Digital Image Processing", Shoko-do, 2002
- [20] S. Mizuta, Y. Du, K. Shioda, "Automatic extraction of human foetal specimen regions from 3D MRM images using region-based contour trees and measurement of body length and volume for large image datasets", Medical Imaging Technology, Vol. 24, No. 4, pp. 238–246, 2006
- [21] X. Zhou, A. Shimizu, J. Hasegawa, J. Toriwaki, "Vision Expert System 3DIMPRESS for Automatic Construction of 3D Image Processing Procedures and Its Performance Evaluation", IEICE Transactions on Information and Systems, Vol. 1, No. 1, pp. 1-16, 1999
- [22] X. R. Shu, T. Hamada, A. Shimizu, J. Hasegawa, J. Toriwaki. "Improvement of the 3D Image Processing Expert System 3D-IMPRESS-Pro and Its Application to the Automatic Construction of Lung Cancer Shadow Detection Procedures", Journal of the Computer-Aided Medical Imaging Society, Vol. 4, No. 1, pp. 1-9, 2000
- [23] T. Nagao, "Evolutionary Image Processing", Shoko-do, 2002
- [24] M. Haru, H. Oshima, H. Suzuki, "Frontiers in Image Diagnosis - PET/CT, 3DCT", Journal of the Japanese Society of Surgery, Vol. 106, No. 11, pp. 677-684, 2005
- [25] S. Nitta, K. Honda, T. Sogaya, "Diagnostic Support System for Automatic Extraction of Tumor Candidate Regions from PET/CT Images", Medical Imaging Technology, Vol. 24, No. 3, pp. 181-190, 2006
- [26] H. Ueki, T. Yamazaki, H. Haneishi, "Proposal of an Absorption Correction Method for Respiratory-Synchronized PET", Medical Imaging Technology, Vol. 24, No. 4, pp. 229-237, 2006
- [27] T. Mizuta, R. Ishikawa, K. Kitamura, A. Ohtani, K. Tanaka, M. Amano, Y. Shimani, H. Tsushima, M. Ibuki, S. Sugawara, Y. Shoji, S. Miura, "Quantitative Evaluation of Scattering Absorption and Absorption Correction in Shimadzu 3DPET", Medical Imaging Technology, Vol. 24, No. 4, pp. 254-261, 2006
- [28] S. Yasuda, M. Ide, "PET and Cancer Screening", Journal of the Japanese Society of Radiological Technology, Vol. 61, No. 6, pp. 759-765, 2005
- [29] H. Chihara, H. Uozumi, "Current Status and Challenges of PET Screening", Journal of the Japanese Society of Radiological Technology, Vol. 61, No. 6, pp. 861-867, 2005

- [30] A. Suzuki, T. Yamaguchi, T. Oka, T. Kawano, K. Shizuki, N. Takahashi, T. Inoue, "Physiological Accumulation in FDG-PET: Differentiation Using SUV", Medical Information Processing Systems Research Society MIPS2004, pp19-20, 2004
- [31] S. Aoki, O. Abe, "Understanding Diffusion MRI", Shuunsha, 2002
- [32] T. Goto, M. Oka, S. Oya, "A Study on the Effectiveness of High b-Values in Head Diffusion-Enhanced Images", Journal of the Japanese Society of Radiological Technology, Vol. 60, No. 8, pp. 1139-1142, 2004
- [33] T. Kiyama, N. Kodama, T. Kaneko, I. Fukumoto, "Study on Diagnostic Techniques for Alzheimer's Disease Using MRI", Transactions of the Institute of Electronics, Information and Communication Engineers, 2002
- [34] T. Kaneko, N. Kodama, T. Kiyama, Y. Kawase, S. Shibukawa, I. Fukumoto, "A Study on an Objective Diagnostic System for Alzheimer's Dementia Based on Automatic Extraction and Feature Analysis of the Corpus Callosum from Mid-Sagittal MRI Images of the Head", Journal of the Japanese Society of Radiological Technology, Vol. 60, No. 2, 2003
- [35] M. Goto, S. Aoki, O. Abe, T. Masumoto, Y. Watanabe, Y. Satake, K. Nishida, K. Ino, K. Yano, Y. Iida, K. Mima, K. Otomo, "Effectiveness of Horizontal Section Images in the VSRAAD System for Early Diagnosis of Alzheimer's Disease", Journal of the Japanese Society of Radiological Technology, Vol. 62, No. 9, 2006
- [36] K. Ishii, "Diagnosing Alzheimer's Disease Using Images" Modern Physician, Vol. 21, No. 4, 2001
- [37] S. Osher, J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations", Journal of Computational Physics, 79(1), 12–49, 1998
- [38] Y. Boykov, M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images", Proceedings of ICCV, 105–112, 2001
- [39] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning", Nature, 521(7553), 436–444, 2015
- [40] H. Park, P. H. Bland, C. R. Meyer, "Construction of an Abdominal Probabilistic Atlas and its Application in Segmentation", IEEE Transactions on Medical Imaging, 22(4), 483–492, 2003
- [41] National cancer institute, Cancer data access system: <https://cdas.cancer.gov/approved-projects/1691/>
- [42] D. R. Aberle, A. M. Adams, C. D. Berg, "Reduced lung-cancer mortality with low-dose computed tomographic screening", New England Journal of Medicine, 365(5), 395–409, 2011
- [43] S. G. Armato, G. McLennan, L. Bidaut, et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI)", A completed reference database of lung nodules on CT scans. Medical Physics, 38(2), 915–931, 2011

- [44] A. Setio, A. Traverso, T. Bel, et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images" The LUNA16 challenge. *Medical Image Analysis*, 42, 1–13, 2017
- [45] B. Van Ginneken, et al., "Computer-aided diagnosis: How to move from the laboratory to the clinic" *Radiology*, 273(2), 318–332, 2015
- [46] W. Shen, M. Zhou, F. Yang, et al., "Multi-scale convolutional neural networks for lung nodule classification" *Information Processing in Medical Imaging*, 9123, 588–599, 2015
- [47] G. Litjens, T. Kooi, B. E. Bejnordi, et al. , "A survey on deep learning in medical image analysis" *Medical Image Analysis*, 42, 60–88, 2017
- [48] Z. Wei et al., "MmWave radar and vision fusion for object detection in autonomous driving: A review", *Sensors*, 2022
- [49] K. Patel et al., "Deep learning-based object classification on automotive radar spectra", *IEEE Radar Conference*, 2019
- [50] Y. Wang et al., "RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization", *IEEE Journal of Selected Topics in Signal Processing*, 2021
- [51] V. K. Le, et al., "Energy efficient hand gesture recognition using event camera and spiking neural networks", *Virtual ITS European Congress*, 2020
- [52] Y. Nakano, et al., "Investigation of performance shift by conversion from convolutional neural network to spiking neural networks", *ITS World Congress*, 2021
- [53] I. Alonso, et al., "EV-SegNet: Semantic segmentation for event-based cameras", *CVPR Workshop Paper*, 2018
- [54] J. Wan, et al., "Event-based pedestrian detection using dynamic vision sensors", *Electronics*, 2021
- [55] D. Singla, et al., "HyNNA: Improved performance for neuromorphic vision sensor-based surveillance using hybrid neural network architecture", *IEEE ISCAS*, 2020
- [56] B. Rueckauer, et al., "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification", *Journal of Frontiers in Neuroscience*, vol. 11, pp. 682, 2017
- [57] L. J. Haeng, et al., "Training deep spiking neural networks using backpropagation", *Journal of Frontiers in Neuroscience*, vol. 10, pp. 508, 2016
- [58] F. Nicolas, et al., "Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules", *Journal of Frontiers in Neuroscience*, vol. 9, pp. 85, 2016
- [59] C. Lee, et al., "Enabling spike-based backpropagation in state-of-the-art deep neural network architectures," *arXiv:1903.06379*, 2019.

- [60] E. O. Neftci, et al., “Surrogate gradient learning in spiking neural networks”, arXiv:1901.09948, 2019
- [61] M. Mozafari, et al., “Spyketorch: Efficient simulation of convolutional spiking neural networks with at most one spike per neuron”, arXiv:1903.02440, 2019
- [62] O. Bichler et al., “N2d2-neural network design and deployment”, Manual available on Github, 2017
- [63] K. He et al., “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification”, in Proceedings of the IEEE international conference on computer vision, pp. 1026–1034, 2015
- [64] P. Filip and K. Andrzej, “Introduction to spiking neural networks: information processing,” Learn. Appl, vol. 71, pp. 409–433, 2011
- [65] D. Heeger, “Poisson model of spike generation”, Handout, University of Stanford, vol. 5, pp. 1–13, 2000
- [66] A. Tavanaei, A. Maida, “Bp-stdp: Approximating backpropagation using spike timing dependent plasticity”, Neurocomputing, vol. 330, pp. 39–47, 2019
- [67] P. U. Diehl et al., “Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing”, International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2015
- [68] H. Bulzomi, S. Memudu, Y. Nakano and J. Martinet “Real-Time Pedestrian Detection at the Edge on a Fully Asynchronous Neuromorphic System”, Conference on Computer Vision and Pattern Recognition workshop on Event-based Vision, 2025
- [69] Li, H., et al. “Spiking-YOLO: Event-based Object Detection with Spiking Neural Networks.”, arXiv:2005.00001, 2020
- [70] S. Zhou, et al., “Real-time Pedestrian Detection with Spiking Neural Networks on Neuromorphic Hardware.”, Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2021
- [71] J. Park, et al., “Energy-efficient Pedestrian Detection using Hybrid Event-driven Spiking Neural Networks.”, IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2022.
- [72] A. Safa, et al., “Fusing Event-based Camera and Radar for Human Detection using Spiking Neural Networks with Continual STDP Learning”, arXiv:2207.00006, 2022
- [73] A. Viale, A. Marchisio, M. Martina, G. Masera, M. Shafique, “LaneSNNs: Spiking Neural Networks for Lane Detection on the Loihi Neuromorphic Processor”, arXiv:2208.02253, 2022
- [74] M. Davies, et al., “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning”, IEEE MICRO, vol. 38, no. 1, pp. 82–99, 2018
- [75] C. E. Mauriello, H. Bulzomi, J. Martinet, Y. Nakano, “End-to-End Neuromorphic Lane Detection”, International Workshop On Advances in Sensors and Interfaces, 2025

- [76] X. Liu, et al., "Towards Energy-efficient Audio-Visual Classification via Multimodal Interactive Spiking Neural Network (MISNet)", ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2025
- [77] W. Li, et al., "Spiking Tucker Fusion Transformer for Audio-Visual Zero-Shot Learning", IEEE Transactions on Image Processing (TIP), 2024
- [78] X. He, et al., "Enhancing Audio-Visual Spiking Neural Networks via Semantic-Alignment and Cross-Modal Residual Learning", arXiv:2502.00003, 2025
- [79] L. Guo, et al., "Transformer-based Spiking Neural Networks for Multimodal Audio-Visual Classification", IEEE Transactions on Cognitive and Developmental Systems (TCDS), 2023
- [80] T. Taunyazov, et al., "Event-Driven Visual-Tactile Sensing and Learning for Robots", Robotics: Science and Systems (RSS), 2020
- [81] J. Yang, et al., "TL-SNN: Event-Driven Visual-Tactile Learning with Temporal and Location Spiking Neurons", SSRN Preprint, 2025.
- [82] A. Safa, et al., "Fusing Event-based Camera and Radar for SLAM Using Spiking Neural Networks with Continual STDP Learning", arXiv:2207.00006, 2022
- [83] B. Vogginger, et al., "Automotive Radar Processing with Spiking Neural Networks", Frontiers in Neuroscience, 2022
- [84] N. Rathi, K. Roy, "STDP-Based Unsupervised Multimodal Learning with Cross-Modal Processing in Spiking Neural Networks", IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI), 2018
- [85] T. Li, et al., "SpikeCLIP: Distilling Multimodal Knowledge into Spiking Neural Networks", arXiv:2310.06488, 2025
- [86] X. He, et al., "Temporal Attention-Guided Spiking Neural Networks for Multimodal Processing", arXiv:2503.00007, 2025

Appendix A

Automatic construction of video image processing using 3D-ACTIT

A.1. Summary

Input moving images into 3D-ACTIT and automatically construct video image processing using a combination of 3D image processing filters. By treating the time axis as the Z axis and video data (frame images) as 3D volume data, we hypothesise that 3D-ACTIT can be used to construct video image processing. In this chapter, we conduct an experiment to automatically construct the process of extracting moving objects from video data using 3D-ACTIT by reproducing a road environment captured by a vehicle-mounted camera using a small robot. We verify whether 3D-ACTIT can construct the desired processing for moving images simply by inputting a set of training images.

A.2. Training dataset

We captured video images simulating a road environment using multiple small robots. Specifically, we mounted cameras on mobile robots and captured images in an environment simulating a road, using mobile robots that simulated oncoming vehicles and vehicles ahead. We created three scenes by changing the background and other elements, and the goal was to extract oncoming vehicles and vehicles ahead from each scene. In other words, we aimed to automatically construct a process for extracting moving objects from scenes captured by on-vehicle cameras. Figure 74 shows the training image set. The training image set includes target images prepared by extracting only the regions of moving objects from two scenes with different backgrounds, as shown in Figure 74. These two sets were input simultaneously. The image processing filter set used is shown in Table 2.

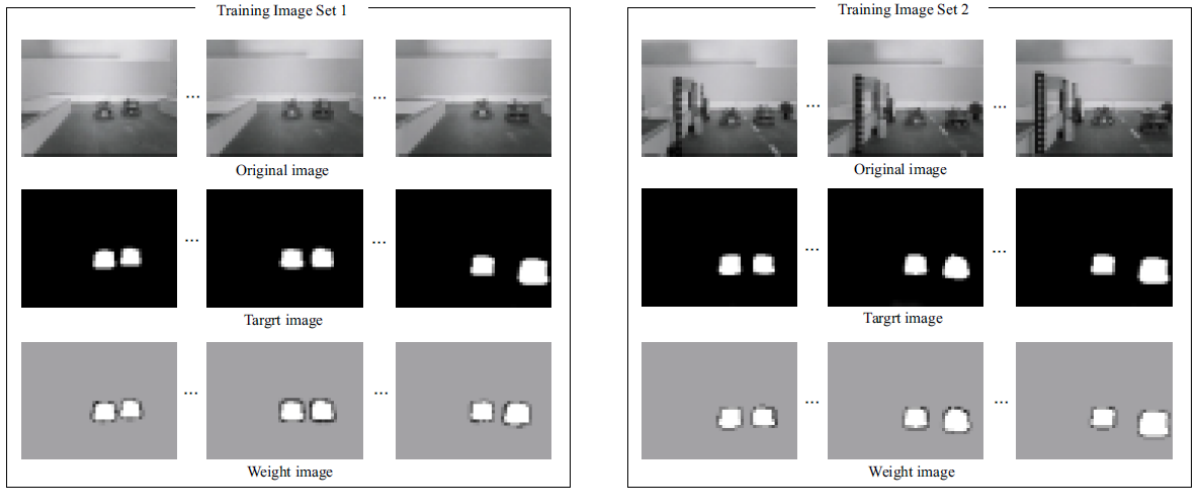


Figure 74 Training image set for moving object extraction

A.3. Moving object extraction from video data

The teacher image set in Figure 74 was input into 3D-ACTIT, and the output images and tree-structured filters are shown in Figure 75. Looking at the output images, it can be seen that they react to objects in the background and contain some noises, however they still extract the regions of moving objects even in different scenes. The results of applying the constructed tree structure to testing images and the results during image conversion are shown in Figure 76. Analysing the tree structure, we generated an image that serves as a mask by using a filter "t" that sets values equal to or below the average tone value of the input original image to white, and a filter "i" that inverts pixel value.

"i" to generate an image that serves as a mask to limit the processing area. By taking the boundary product "b" between this mask image and the image transformed by the other subtree, only the edges of moving objects are extracted. In the latter half of the processing, it can be seen that the target image is approximated using neighbourhood processing filters such as the average value filter "-" and the maximum value filter "M", as well as the binarization filter "n".

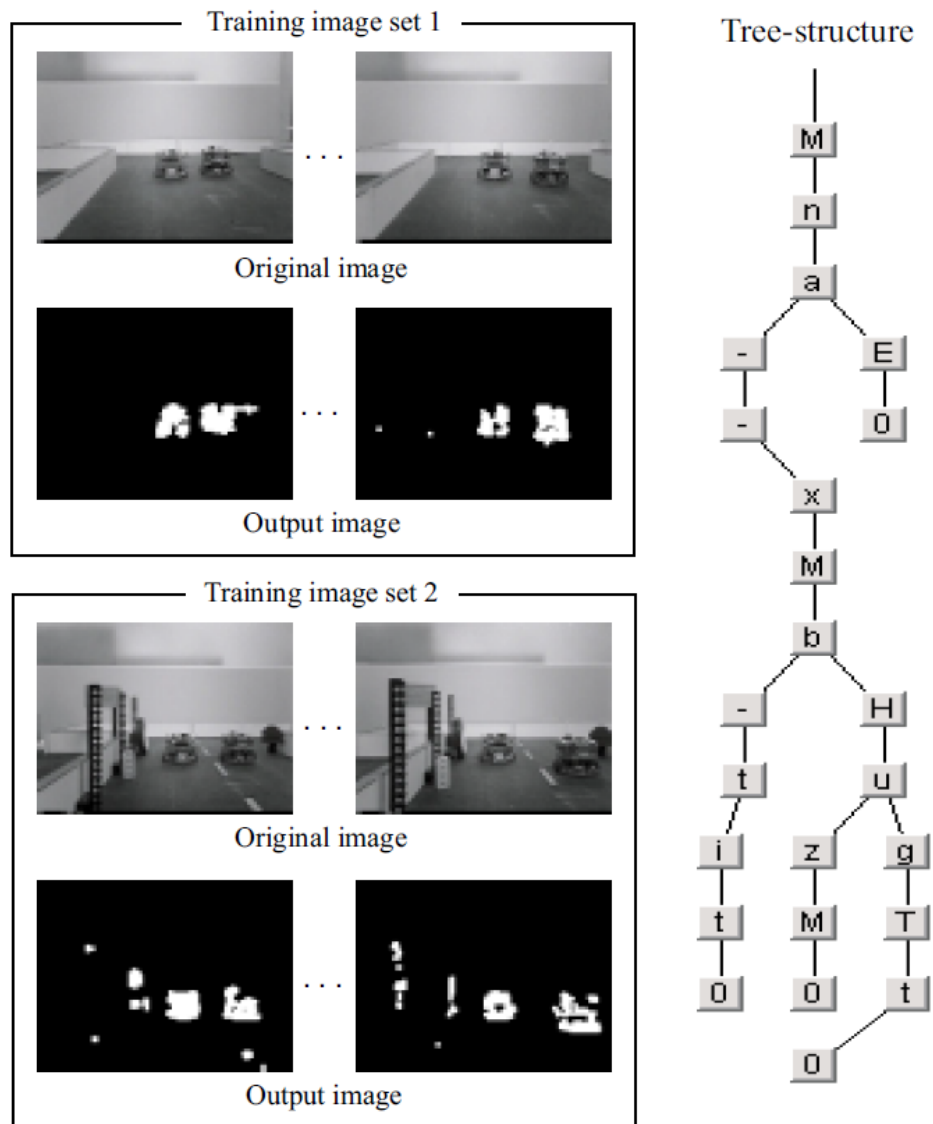


Figure 75 Results on training data for moving object extraction

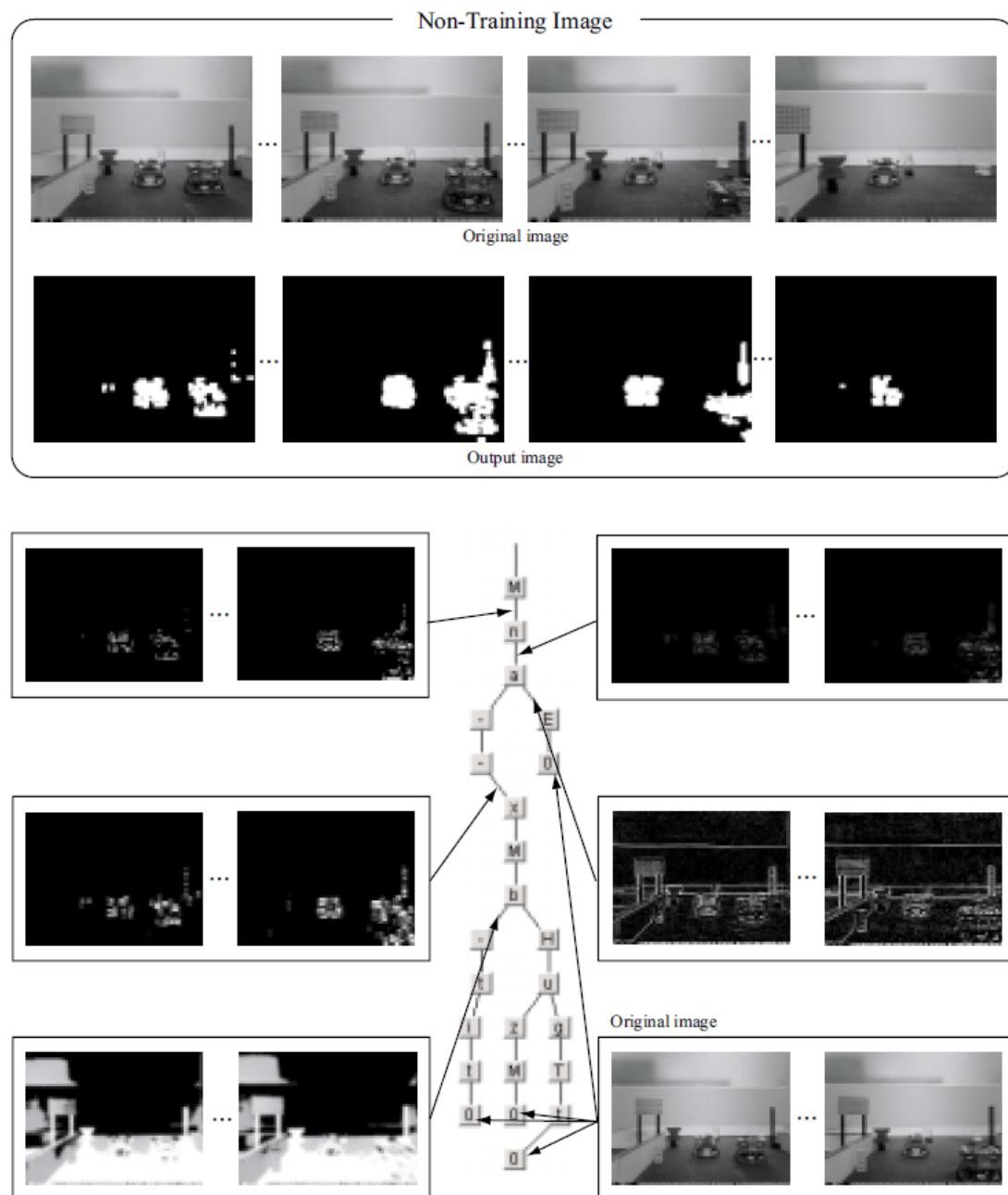


Figure 76 Results on testing data