SEMANTIC SEGMENTATION OF HANDS IN MULTIMODAL IMAGES: A NEW REGION-BASED CNN APPROACH

Akila Pemasiri\(^1\), David Ahmed-Aristizabal\(^1\), Kien Nguyen\(^1\), Sridha Sridharan\(^1\)
Sasha Dionisio\(^2\), Clinton Fookes\(^1\)

\(^1\) Image and Video Research Laboratory, SAIVT, Queensland University of Technology, Australia.
\(^2\) Department of Mater Advanced Epilepsy Unit, Mater Centre for Neurosciences, Australia.

ABSTRACT

Segmentation of body parts is a critical but challenging stage in the medical imaging processing pipeline due to the anatomical complexity. Recent advances in deep learning have successfully dealt with this complexity in visible images. However, the efficacy of these segmentation techniques for other modalities of interest such as X-ray images is not known. We propose a unified semantic segmentation approach for body parts for both X-ray and visible images which can be concurrently applied to both modalities. Unifying the two modalities in a single model not only reduces the number of parameters, but also improves the accuracy by enabling end-to-end training and inference. Quantitative results are validated on two clinical applications: (1) a static analysis of hand segmentation in visible and X-ray images; and (2) a dynamic analysis which quantifies and classifies epileptic seizures from clinical manifestations of visible hand and finger motions. The proposed model is a potential stepping stone towards developing more robust automated systems that support the assessment of medical conditions based on multimodal information.

Index Terms— Deep learning, X-ray images, Epilepsy, Quantitative motion analysis.

1. INTRODUCTION

Medical image processing in the clinical setting is a growing field of computer vision as it reinforces the benefits of noncontact systems for objective assessment of a patient’s condition. As an early stage of the processing pipeline, segmentation is a critical but also a challenging task considering the complexity of the anatomical structure evaluated and the nature of the imaging devices [1]. Recent attention has been focused on semantic segmentation, which is an evolved version of the traditional segmentation task. Compared with the rectangular bounding boxes of the objects output from the traditional segmentation, semantic segmentation attempts to partition the image into semantically meaningful parts with more fine-grained masks rather than rectangular ones. Semantic segmentation can be very useful for medical imaging analysis of the human body due to a more accurate anatomical localization of the region of interest [2].

Deep learning and segmentation have been utilized for the analysis of images such as X-ray, computed tomography scan (CT) and magnetic resonance (MR) for different applications including anatomical structure identification [3], tumour localization [4] and bone fracture detection [5]. Recent machine learning advances have also enabled the development of methods to support medical diagnosis from video recordings of visible images including facial assessment such as facial paralysis [6], patient monitoring, and epilepsy evaluation [7].

Despite the success of deep learning and region-based approaches in clinical scenarios with visible images [8], these approaches still struggle to provide interpretable evidence for their decisions, by precisely locating the body parts under evaluation. For example, when analysing the features from visible images the existing systems are unable to clearly locate and refine the boundaries of the human body parts, and as a result motions of other objects in the environment are included in analyses, which can adversely affect the results. In this scenario, utilising semantic segmentation can assist in accurately locating the region of interest.

Furthermore, in medical diagnosis processes there are situations where multiple imaging modalities (e.g. standard digital images, X-ray images, thermal images) are used in combination [9, 10, 11]. When adopting computer vision techniques to support physicians, the use of different imaging modalities in automatic diagnosis necessitates the development of image processing techniques that are invariant to the image modalities and concurrently capture salient information in all modalities. However, multimodal image processing tasks present challenges that are inherent to multimodal images, including visual dissimilarities such as different textures and different levels of information [12].

In this paper, we address the above challenges and present a method to establish object boundaries and pixel-wise semantic object segmentation of human body parts regardless of the modality, whether it is an X-ray or a visible image. The semantic segmentation method we propose has the capacity to enhance the diagnosis by selecting the most discriminative features of the anatomical parts under consideration and the multimodal segmentation that we propose has the potential to enhance the applicability of computer vision techniques in common diagnosis processes. We validate our segmentation approach by investigating the detection performance on the public X-ray image MURA dataset [13], which contains X-ray images of the elbow, forearm, hand, humerus, shoulder and wrist. This paper also seeks to ascertain if incorporating a semantic segmentation phase will enhance the classification accuracy of epilepsy types using hand motions captured on our video dataset collected on an Australian hospital. To the best of our knowledge, this is the first work that considers modality-invariant semantic segmentation for visible and X-ray images.

The contributions of our work are summarized as follows:

1. We introduce a novel multimodal semantic segmentation approach for assistive medical diagnosis based on region-based deep learning and demonstrate the approach using hand images.

2. We present the first semantic segmentation approach of X-ray images that forms the basis for further research to capture deep features and assess medical conditions such as bone densitometry [14], rheumatoid arthritis identification [15] and bone fractures [5].

3. We propose an enhanced system to classify epilepsy types considering hand and finger motions following the study pro-
posed in [8]. We incorporate the multimodal segmentation methodology to improve the classification performance under challenging imaging conditions typical of an epilepsy monitoring unit.

2. METHOD

In this paper, we present an approach that is capable of performing semantic segmentation for visible and X-ray images of hands using one model, which can also be extended for identifying semantic segmentation for other body parts and modalities as well. We demonstrate the benefits of the proposed architecture illustrated in Fig. 1, which is based on region-based Convolution Neural Network (CNN) frameworks, by providing two applications under extremely different imagining scenarios: (1) Static analysis of X-ray images of human upper extremity; and (2) Dynamic analysis of visible images captured on patients with epilepsy for assessment of semiology. In contrast to the static analysis where an independent instance of a patient was subjected for the inference, in the dynamic analysis we investigate the effect of establishing accurate segmentation in a sequence of image where the analysis is based on both temporal and spatial features.

Details of the model architecture and strategy for each clinical analysis are described in the following subsections.

2.1. Semantic segmentation architecture

While Mask R-CNN [16] is the most widely used semantic segmentation method, its performance is diminished when considering multimodal images such as X-ray and visible images under natural clinical settings. In the existing CNN approaches which are for semantic segmentation, the layers are arranged in a cascaded manner where the successive layers use the outputs from previous layers as the input. By observing the outputs of the Mask R-CNN, it was identified that not all the layers in the network generate features for different modalities as depicted Fig. 2. Inspired by this observation, we define a new architecture displayed in Fig. 1, where the layers that generate discerning features for both modalities are fused as a single layer. This layer is then subjected to class predictions, Region of Interest (RoI) identification and mask identification tasks, using a multi-task loss function.

For the multimodal segmentation network, we trained our network with a dataset which contains X-ray and visible images. For the X-ray image dataset, we used the MURA dataset [13] which contains X-ray images of the human body parts. In this dataset, we generated manually annotated mask ground-truth. The dataset of visible images contains images captured from 30 subjects, under different conditions including changes in background, resolution, illumination, as well as occlusion by accessories such as wrist watches, jackets and bangles. We captured a total of 1400 images which include annotations of 817 hands, 908 wrists, 1023 forearms, 986 elbows, 807 humerus and 722 annotations for the shoulder.

Section 2.2 and Section 2.3 provide a detailed description of using this architecture in different settings.

2.2. Static analysis: segmentation in X-ray images

2.2.1. Overview

The existing approaches for segmentation in X-ray images have focused on identifying the components of human musculoskeletal system (i.e. identifying bones or muscles) as depicted in Fig. 3(b) [17, 18]. In contrast to that, by using the architecture described in Section 2.1, more semantically meaningful segmentation can be obtained (Fig. 3(c)). This enables better localisation in X-ray images for the identification of medical conditions such as congenital deformities.

2.2.2. Dataset and analysis

For the X-ray image investigation, we used the same dataset that is mentioned on Section 2.1. It should be noted that the research reported in this paper has targeted only on human upper extremity due to the unavailability of X-ray image datasets for other body parts; however, the method proposed, can be easily applied to other body parts as well. Some instances of semantic segmentations obtained by our method are depicted in Fig 4.

2.3. Dynamic analysis: hand motion analysis in epilepsy

2.3.1. Overview

During epileptic seizures, a myriad of clinical manifestations known as semiology may occur. Semiology provides localizing information of the brain networks affected, enabling the progression to successful surgery [19]. Certain hand and finger motions such as hand dystonia, tapping, snapping, or thumb adduction [19, 20] are useful to distinguish between epilepsy types. Nevertheless, the study of these signs
We expand the detected patient bounding box with an offset of 20% pre-trained Mask-RCNN weights, trained on the COCO dataset [21]. Semantic segmentation is dependent on the experience of the clinician and the interpretation can at times differ from physician to physician, and from case to case.

Automated assessment of semiology has shown promise in extracting quantifiable information from behavior to help define epileptic seizures [7], which would enable more objective information gathering from epileptic patients. A novel method to characterize ictal semiology from hand and finger motions and to differentiate between patients with mesial temporal (MTLE) and patients with extratemporal lobe epilepsy (ETLE) was introduced in [8]. This approach is a region-based method which extracts spatial features from a detected hand-bounding box, which are subsequently fused to capture the temporal change of the whole hand. Despite its success in quantifying these clinical manifestations, the region-based approach is still affected by capturing information that is irrelevant to the hand semiology such as motions in bedding and monitoring equipment. Additionally, by extracting features from the complete hand-bounding box detected, there are events where the information of both hands are overlapped. This situation is shown in Fig. 6. Therefore, we argue that by including an invariant segmentation phase to the system, the analysis of this type of semiology can be improved.

2.3.2. Dataset

We acquired data as part of the routine Video-EEG (scalp electroencephalography) and Video-SEEG (stereo-electro-encephalography) monitoring protocol from patients with epilepsy at the Mater Hospital in Brisbane. Assessment and categorization of all recorded seizures was conducted by clinical experts to ensure the validity of the supervised deep learning model. A total of 10 video clips were recorded, consisting of 7 videos from patients with MTLE and 3 videos from patients with ETLE. To quantify semiological features, the inputs of the system are short video sequences rather than a whole video, such that we obtain more data to train the system. We define a sequence as five consecutive frames. Input video is captured at 25 frames per second, and we downsample footage by a factor of 5 to extract more pronounced motion and better recognize seizures.

2.3.3. Hand semiology quantification strategy

The extraction of hand features is based on [8]. Following their study, we implement the proposed region-based approach coupled with our segmentation phase. A block diagram of the proposed system for the hand analysis is displayed in Fig. 5.

To extract features related to the patients behavior, we first define the region of interest that contains the patient. We perform patient boundary detection using the Mask-RCNN architecture [16]. We use pre-trained Mask-RCNN weights, trained on the COCO dataset [21]. We expand the detected patient bounding box with an offset of 20% of the total width on each side to avoid the extremities of the patient being located outside of this boundary due to movements during a seizure. Then, a hand detection phase is performed to improve the segmentation performance. To this end, we chose to follow the architecture in [22], where the hand bounding box is detected based on the body key-point prediction [23]. The architecture is evaluated on two challenging hand databases: Oxford hand dataset [24] and VIVA Challenge [25].

Once the semantic segmentation phase is executed, spatial features are extracted from the body part number 6 (hand) and the last fully connected layer with a dimension of [1,4096], as the output of the layer in the network has 4,096 units. Each sequence of features of the movements has a dimensionality of [5,4096], capturing five frames, each with 4,096 features. We further capture the dynamic variations of the hand in sequences. The spatial representation of each hand is fed into an LSTM network, which is capable of learning long-term dependencies present in the sequential data [26]. We adopt a configuration with two stacked LSTM layers, each with 128 memory cells. More complex architectures do not show significant performance gain. The outputs of each LSTM are concatenated into a single densely connected layer with a sigmoid activation function to make a single prediction for every sequence for each hand. The final hand accuracy is defined as the average performance between the left and right hand classification of the ictal movement detected.

We train the LSTM networks by optimizing the binary cross entropy loss using the Adam optimizer [27] with a learning factor of $10^{-3}$, and the first and second moment decay rates of 0.9 and 0.999, respectively. We adopt a batch size of 32 and dropout with a probability of 0.35. We perform the training using 200 epochs with the default initialization parameters (the weights of the LSTM hidden units) from Keras [28].

3. EVALUATION

3.1. Experimental setup of X-ray images

To evaluate our method, we compared our method with Mask R-CNN [16], which is the state-of-the-art method for semantic segmentation. We trained the Mask R-CNN and our architecture on the same dataset, and tested on a set of X-ray images which were not used for the training. We adopt as evaluation matrix the average precision (AP), which is calculated using the Intersection over Union (IoU) based on a defined threshold,

$$IoU = \frac{\text{Area} (b_p \cap b_i)}{\text{Area} (b_p \cup b_i)}$$

In Equation 1, $b_p$ is the bounding box prediction or the mask prediction, and $b_i$ is the corresponding ground-truth value. In this eval-
Fig. 5: Framework of the region-based method enhanced by incorporating a segmentation phase to distinguish hand semiology. Detection of the patient and hands are performed to improve the accuracy of the segmentation. Spatial information is extracted from the hidden layer activation of our semantic segmentation architecture. The spatiotemporal relation between frames for each hand is analyzed using a LSTM to predict the class of the sequence.

Table 1: Average precision on X-ray images for mask prediction and bounding box predictions, where the models have been trained on multimodal dataset.

| Approach       | Backbone      | Boundingbox Prediction Accuracy | Mask Prediction Accuracy |
|----------------|---------------|--------------------------------|--------------------------|
| Mask R-CNN     | Resnet-50     | 42.05                          | 37.7                     |
| Mask R-CNN     | Resnet-101    | 49.34                          | 44.3                     |
| Mask R-CNN     | ResNeXt-50    | 48.63                          | 42.1                     |
| Mask R-CNN     | ResNeXt-101   | 50.34                          | 44.8                     |
| Ours           | Resnet-50     | 57.28                          | 53.4                     |
| Ours           | Resnet-101    | 61.70                          | 56.4                     |
| Ours           | ResNeXt-50    | 62.49                          | 55.8                     |
| Ours           | ResNeXt-101   | 63.58                          | 56.9                     |

3.2. Experimental setup of visual images in epilepsy

We adopted a k-fold cross-validation [31], with k set to 10, in order to validate the generalization of our automated strategy of hand motion quantification. For this evaluation, the sequences of all patients of the same class are split into 70% for training, 15% for validation and 15% for testing. In this scenario, the sequences used for validation and testing are completely separate to those used for training, but it is possible to have sequences from the same patient in each set. The validation and test accuracy of the framework is computed as the average performance of each fold. A leave-one-patient-out cross-validation scheme will be considered in future experiments as we incorporate more participants with the isolated hand motions.

Table 2: 10-Fold Cross-Validation. Comparison of performance of a region-based approach with and without a segmentation phase in the system.

| Approach               | Validation Accuracy(%) | Test Accuracy(%) |
|------------------------|-------------------------|------------------|
| Region-based [8]       | 90%                     | 75%              |
| Region-based(ours)     | 93.4%                   | 82.7%            |

3.3. Experimental results and discussion

3.3.1. Segmentation in X-ray images

We trained the Mask R-CNN model and our model using the same dataset. Table 1 displays the average precision values of both models in the test set. From these results, it can be seen that for all the backbone architectures our method has surpassed Mask R-CNN, by reaping the benefits of utilizing discerning feature outputs.

3.3.2. Epilepsy quantification and classification

The cross-validation accuracy is reported in Table 2. The deep framework coupled with the segmentation phase was capable of achieving an average test accuracy of 82.7%. This represents an improvement on the hand performance analysis compared with the strategy of extracting spatial features from the bounding box of the hand detected [8].

4. CONCLUSION

This paper presents a novel modality-invariant semantic segmentation to analyse multimodal images. We demonstrate the robustness of our system by providing the first clinical application of hand segmentation in X-ray images and an improved hand motion analysis in epileptic patients. Experimental results confirm that the segmentation approach based on region-based CNN architectures provides additional capacity to model the variety of clinical images and is a promising option for assistive medical diagnosis, where currently automated approaches are adversely affected by irrelevant features and unconstrained conditions encountered in natural clinical settings. Having a single model for multimodal image segmentation, will not only enable the seamless segmentation, but will also reduce the manual involvement by the medical professionals in the process.

Acknowledgment

The research presented in this paper was supported by an Australian Research Council (ARC) grant DP170100632.
5. REFERENCES

[1] Rutvi Shah and Priyanka Sharma, “Bone segmentation from x-ray images: Challenges and techniques,” in Information Systems Design and Intelligent Applications, pp. 853–862. Springer, 2018.

[2] Cristina Stolojescu-CriSan and Ştefan Holban, “A comparison of x-ray image segmentation techniques,” Advances in Electrical and Computer Engineering Engineering, vol. 13, no. 3, 2013.

[3] Xiangrong Zhou, Ryosuke Takayama, Song Wang, Xinxin Zhou, Takeshi Hara, and Hiroshi Fujita, “Automated segmentation of 3D anatomical structures on CT images by using a deep convolutional network based on end-to-end learning approach,” in Medical Imaging 2017: Image Processing, 2017, vol. 10133, p. 1013324.

[4] Hui Cui, Xiuying Wang, and Dagan Feng, “Automated localization and segmentation of lung tumor from PET-CT thorax volumes based on image feature analysis,” in EMBC. IEEE, 2012, pp. 5384–5387.

[5] Martin Donnelly and Greg Knowles, “Computer aided long bone fracture detection...” in ISSPA, 2005, pp. 175–178.

[6] Ting Wang, Shu Zhang, Junyu Dong, Lian Liu, and Hui Yu, “Automatic evaluation of the degree of facial nerve paralysis,” Multimedia Tools and Applications, vol. 75, no. 19, pp. 11893–11908, 2016.

[7] David Ahmetd-Aristizabal, Clinton Fookes, Sasha Dionisio, Kien Nguyen, João Paulo S Cunha, and Sridha Sridharan, “Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey,” Epilepsia, vol. 58, no. 11, pp. 1817–1831, 2017.

[8] David Ahmetd-Aristizabal, Clinton Fookes, Simon Denman, Kien Nguyen, Sridha Sridharan, and Sasha Dionisio, “A hierarchical multi-modal system for motion analysis in epileptic patients,” Epilepsy & Behavior, vol. 87, pp. 46–58, 2018.

[9] Bo-Lin Jian, Chieh-Li Chen, Wen-Lin Chu, and Min-Wei Huang, “The facial expression of schizophrenic patients applied with infrared thermal facial image sequence,” BMC psychiatry, vol. 17, no. 1, pp. 229, 2017.

[10] T Nakamura, M Sato, and H Kajimoto, “semi-automatic scoring method for torticollis by using kinect: 328,” Movement Disorders, vol. 28, pp. 120, 2013.

[11] Robin J Hennessy, Patrizia A Baldwin, David J Browne, Anthony Kinsella, and John L Waddington, “Three-dimensional laser surface imaging and geometric morphometrics resolve frontonasal dysmorphology in schizophrenia,” Biological psychiatry, vol. 61, no. 10, pp. 1187–1194, 2007.

[12] Atousa Torabi and Guillaume-Alexandre Bilodeau, “A Iss-based registration of stereo thermal–visible videos of multiple people using belief propagation,” Computer Vision and Image Understanding, vol. 117, no. 12, pp. 1736–1747, 2013.

[13] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al., “Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs,” arXiv preprint arXiv:1712.06957, 2017.

[14] M Vinoth and B Jayalakshmi, “Bone mineral density estimation using digital x-ray images for detection of rheumatoid arthritis,” International Journal of Pharma and Bio Sciences, vol. 5, no. 3, pp. 104–121, 2014.

[15] Stephen J Wilkinson, Amy Naylor, and KL Goh, “Digital image analysis protocol for determining the radiocarpal joint space in the rheumatoid arthritic wrist,” Computers in biology and medicine, vol. 89, pp. 127–134, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, “Mask r-cnn,” in ICCV. IEEE, 2017, pp. 2980–2988.

[17] GK Manos, AY Cairns, IW Ricks, and D Sinclair, “Segmenting radiographs of the hand and wrist,” Computer Methods and Programs in Biomedicine, vol. 43, no. 3-4, pp. 227–237, 1994.

[18] Mohammed Ameer Ali, Laurence S Dooley, and Gour C Karmakar, “Object based image segmentation using fuzzy clustering,” in ICASSP. IEEE, 2006, vol. 2, pp. II–II.

[19] Soheyl Noachtar and Astrid S Peters, “Semiology of epileptic seizures: a critical review,” Epilepsy & Behavior, vol. 15, no. 1, pp. 2–9, 2009.

[20] Barbara C Jobst, Adrian M Siegel, Vijay M Thadani, David W Roberts, Harker C Rhodes, and Peter D Williamson, “Intractable seizures of frontal lobe origin: clinical characteristics, localizing signs, and results of surgery,” Epilepsia, vol. 41, no. 9, pp. 1139–1152, 2000.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in ECCV. 2014, pp. 740–755.

[22] Tom Simon, Hanbyul Joo, Iain A Matthews, and Yaser Sheikh, “Hand keypoint detection in single images using multi-view bootstrapping,” in CVPR, 2017, vol. 1, p. 2.

[23] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in CVPR, 2017.

[24] Arpit Mittal, Andrew Zisserman, and Philip HS Torr, “Hand detection using multiple proposals.,” in BMVC, 2011, pp. 1–11.

[25] Nikhil Das, Eshed Ohn-Bar, and Mohan M Trivedi, “On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics,” in ITSC, 2015, pp. 2953–2958.

[26] Klaus Greff, Rupesh K Srivastava, Jan Koutnık, Bas R Stengel, and Jürgen Schmidhuber, “Lstm: A search space odyssey,” IEEE transactions on neural networks and learning systems, 2017.

[27] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[28] François Chollet et al., “Keras,” 2017.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[30] Soheyl Noachtar and Astrid S Peters, “Semiology of epileptic seizures: a critical review,” Epilepsy & Behavior, vol. 15, no. 1, pp. 2–9, 2009.