8
Newspaper Segmentation

In this chapter, we propose an algorithm for segmentation of newspaper document images into regions containing text "articles". We target this work towards segmentation of noisy degraded old newspaper images. These are Bangla newspaper images taken from the period 1937-1980. We develop a convolutional neural network with skip connections. The network has two branches, one for image reconstruction (used as an auxiliary task) and the other for the main task of binary segmentation. We also perform denoising of newspaper images using existing PixelRL approach and show the advantage of denoising on the segmentation efficiency.

This chapter is organised in the following manner: Section 8.2 discusses the related work on article segmentation in document images. Section 8.3 formally presents our work elaborating the proposed methodology and architectural details. Section 8.4 describes the experimental setup and dataset. Section 8.5 discusses the obtained result and compares it with existing methods. Finally, section 8.6 concludes this chapter.

8.1 INTRODUCTION

Digitization of old documents is performed to preserve the knowledge present in them for the future generation. There have been various efforts to digitize the documents belonging to different domains like, ancient manuscripts, rare books in different Indian languages, etc., (Ganesan et al. [2015]; Khan and Ali [2017]). The main goal of digitization of document images is to access the content level information present in it. Here, we are considering old Bangla newspapers belonging to the period 1937-1980. Segmentation of newspaper images into its semantic components is an essential step for content-based information retrieval.

Article segmentation forms a preliminary step towards digitization of newspaper images. There are variety of components present in a newspaper page. These components include, news articles, advertisements, figures, weather report, comic section, etc. Classical print media (e.g. journals, magazines, and newspapers) have two essential components: text and graphics. Graphics involves figures, images, symbolic representation, sketches, etc. Segmenting text belonging to articles in a newspaper is a challenging task due to variation in style and complex layout.

The Bangla newspaper images considered in our work belong to the era of 1937-1980. These newspaper images are degraded due to wear and tear over the period of time. Also, these scanned images are noisy with background texture, which adds to its complexity. Figure 8.1 shows a sample Bangla newspaper and its corresponding annotated image.

We develop a deep model for the task of article segmentation in noisy old Bangla newspaper images. Our network consists of three main parts: feature extraction, reconstruction and segmentation. The reconstruction branch ensures efficient feature extraction. The Bangla newspaper images in the dataset are degraded and noisy, which pose severe challenge to identify the separating boundaries between articles. To solve this issue, we have considered clean hindi newspaper images for training our network. We train our network on both clean and noisy images



Figure 8.1 : Sample Bangla newspaper and its corresponding annotated image

and then fine tune it for noisy Bangla images.

8.2 RELATED WORK

Semantic Segmentation of images is a widely researched area. One popular model originally proposed for medical images is U-Net (Yin et al. [2022]). The U-Net is a convolutional neural network (CNN) designed to address image segmentation problems. It consists of an encoder and decoder network, with the encoder network creating a feature map of the input picture to identify the content. The effectiveness of U-Net lies in its use of skip connections and decoder networks. The decoder network determines the location of items in the picture using skip connections to create a segmentation mask. The model design uses contextual feature information from encoder blocks for the segmentation map.

Though semantic segmentation of document images is well explored, segmentation approaches for newspaper articles are relatively few (Furmaniak [2007]; Palfray et al. [2012]; Sfikas et al. [2016]; Chathuranga and Ranathunga [2017]),(He et al. [2017]; Yang et al. [2017]; Meier et al. [2017]).

Unsupervised segmentation of digitized newspaper has been proposed in (Furmaniak [2007]). Their methodology is composed for three steps. Firstly, the document page is processed through an OCR engine to obtain blocks corresponding to paragraphs. In the next step, similar neighbouring block pairs that are likely to be the part of same article are grouped together using text similarity measure. Further, to refine the segmentation maps, geometric rules are applied

which output the final segmented newspaper image.

Palfray et al. (Palfray et al. [2012]) extract the metadata in the images of the digitized newspapers taken from "Journal de Rouen", a regional French newspaper. The extracted logical structure of the newspaper image produces informative units: the articles. Each image is labelled at the pixel level using a Conditional Random Field model. Then a set of rules are defined based on horizontal and vertical separators, titles and text lines to define the concept of article within a newspaper.

Article regions are produced as a result of merging sub-article level content and title regions. Using this concept, a Bayesian Gaussian mixture model approach is proposed in (Sfikas et al. [2016]) for segmentation of a newspaper image into articles. Connected components are extracted and input to the Fully Bayesian Gaussian mixture model (FBGMM) cluster the input into sub-article components, i.e., region types and title groups. The Bayesian model is conditioned on a prior distribution over region features, aiding classification into titles and content. Usage of a Dirichlet prior facilitates the correct estimate of the number of title and article regions.

A procedural approach is proposed in (Chathuranga and Ranathunga [2017]) for article segmentation in newspaper images. They used 12 basic functions which are implemented sequentially for article segmentation. These functions include, graphics extraction, background line detection, column separation, page left/right margin detection, estimating a dynamic threshold coefficient, text extraction, foreground line detection, foreground and background line prolonging, page header/footer detection, filling parallel gaps between lines, recursive flood filling and segmentation and smoothing segments.

Multi-scale and multi-task Fully Convolution Network is proposed in (He et al. [2017]) for semantic page segmentation along with table detection. The network uses state-of-the-art deep learning techniques to segment documents images into three classes viz., blocks, tables, and figures. Conditional Random Field (CRF) is used for output refinement. Their network consist of two branches for two tasks: class label prediction for each pixel, and contour detection for each class. In contour detection, it predicts the boundaries for each class. After obtaining these two prediction maps, CRF is jointly trained on the top of these two branches to produce the final refined output. Their results are presented on Various Doc Dataset which is a collection of 130 pages. The images of this dataset are taken from various document types belonging to magazines, research article and not specific to newspaper.

A Fully Convolution Network (FCN) was proposed in (Meier et al. [2017]) to perform article segmentation in newspaper images. It is built on three logical parts, namely, feature extraction, upscaling, refinement. Feature extraction performs standard CNN operations (convolution and max-pooling) to extract feature maps. It is followed by upscaling and segmentation where, element-wise sum and transpose convolution is applied to perform the task of segmentation. Further refinement is applied to correct the artefacts and sharpen the boundaries of the segmented rectangular regions. The input to this network are clean images of Swiss Newspaper of the year 2016 images along with its OCR pre-processed image. OCR pre-processed image indicates the regions where text and image are found, also it contains horizontal and vertical lines present in newspaper page.

In another related work, the Multimodal Fully Convolution Network (MFCN) proposed in (Yang et al. [2017]) is demonstrated to work for magazines and technical articles. It segments the page into various classes including, figure, table, section heading, caption,list and paragraph. Their model consist of four parts: an encoder, decoder, auxiliary decoder and bridge. An encoder branch is responsible for feature extraction process. Decoder branch is task specific and is used for semantic segmentation task. An auxiliary decoder is used for reconstruction of input, for

the purpose of facilitating efficient feature extraction. An auxiliary decoder is used only during training. A bridge merges the textual and visual representations. The document images considered in this work are synthetically generated clean images.

All the above described research work, segment the clean document images into semantic components. Clean documents exhibits clearly demarcated layouts and articles which facilitates better performance. Using noisy, degraded, old newspaper images pose the real challenge. Our model is inspired from (Yang et al. [2017]) and is explained in the next section.

8.3 NETWORK ARCHITECTURE FOR NEWSPAPER SEGMENTATION

The proposed model consist of three components: feature encoder, reconstruction and segmentation branch. The feature encoder branch is responsible for feature extraction process. The reconstruction branch provide an auxiliary target to facilitates efficient feature extraction. It is used only during training. Segmentation branch is task specific and is used for article segmentation task.

Feature encoder consists of two types of operations. The input image is first passed through standard convolution operation followed by the atrous convolution. Starting layer of the network comprises convolution layer which learns the primitive features. The following layers of atrous convolution layer is also known as dilated convolution.

Dilated convolution was proposed in (Yu and Koltun [2015]) for multi-scale context aggregation. In this type of convolution , the spacing between kernel values is increased and weights are placed far away at given intervals. These intervals are specified by the dilation factor. This delivers the expanded receptive field at the same computational cost without losing the resolution. The use of dilated convolution facilitates the segmentation of small characters present in noisy background (Hamaguchi et al. [2018]).

The feature encoding is the input to two identical branches but with different tasks: segmentation and reconstruction. Segmentation branch is responsible for article segmentation while reconstruction branch is only used during training. These branches contain the convolution layer followed by the upsampling layer. Alternate skip connections are used to allow interaction between low-level and high-level features obtained from dilated convolution. Low-level representations, limited by the local receptive field, are not aware of object-level semantic information. High-level features are not necessarily aligned consistently with object boundaries as CNN models are invariant to translation. These alternate skip connections provide links between the feature encoder and both the segmentation and reconstruction branches.

Figure 8.2 shows the architecture of our proposed network. The given input image is scaled down to $(512, 512)$ dimension and is input to our network. Feature encoder contains two types of layers, namely, convolution layer and dilated convolution. The convolution layer computes the primitive features within the local receptive field. The dilated convolution layer considers a broad receptive field, thus computing more advanced features. The area considered in dilated convolution depends on its dilation rate. There are 5 convolution layers with 64 filters each. Each convolution layer is followed by a Batch Normalization and ReLU activation layer. These 5 layers are followed by Dropout and Max Pooling layers. The primitive features computed from the convolution layer feed into the Dilated Convolution layer. Similar to convolution layer, each dilated convolution is also followed by a Batch Normalization and ReLU activation layer. There are 4 dilation blocks considered in our architecture. Each block has 5 dilation layers, each with a dilation rate $d = 1, 2, 4, 8, 16$ respectively. After each dilation block there is a Dropout and Max Pooling layer.

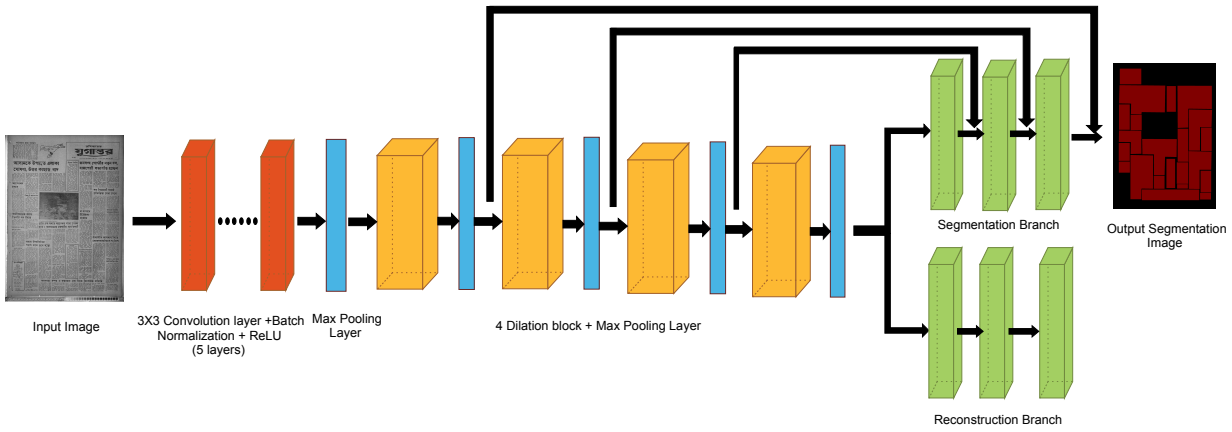


Figure 8.2 : Proposed Architecture

The output of the feature encoding layer is fed to task specific branches dealing with reconstruction and segmentation. Each of these branches has 3 sets of Convolution layers + Batch Normalization + ReLU activation layer + Upsampling layer. Alternative skip connection is implemented between the feature encoder layer and both the reconstruction and segmentation layer.

We train our network on an Nvidia DCX2 50 Epochs, with Adadelta gradient descent algorithm with a learning rate of 0.01. The Hough transform is applied on output images to refine them and to obtain polygons from the segmentation mask.

8.3.1 Dataset

The Bangla newspaper images in our dataset are scanned images available from NDLI (National Digital library of India). These images belong to the years 1937-1980 and are available as gray level images. These images are noisy and degraded with low resolution. These newspapers were printed in gray scale and use limited graphics. The background fabric on which it was printed adds texture to the images which further adds to its complexity. The content present in these images are news articles, comics and advertisements.

For training the segmentation model we collected a dataset of Hindi newspaper images belonging to the modern printing era. We used the e-papers present over the news websites. These are clean images, rich in color. It uses abundance of graphics with pictures embedded in almost all the articles. here is diverse content present in these newspaper images, such as news articles, figures, captions, weather reports, comic section, puzzles, advertisements, etc. We took 100 images of each type for training the model. These images are annotated using the labelme annotation toolbox (Wada [2016]). We used data augmentation to increase the number of images for training. Table 8.1 and 8.2 shows some sample Bangla and Hindi newspaper images taken from the dataset.

8.4 EXPERIMENTS

We experiment with two variants of our model namely, M-single and M-proposed. It is trained on Nvidia DGX-2. We started our work with only feature encoder and segmentation branch, which we termed as M-seg. M-seg-rec is our proposed model which contains feature

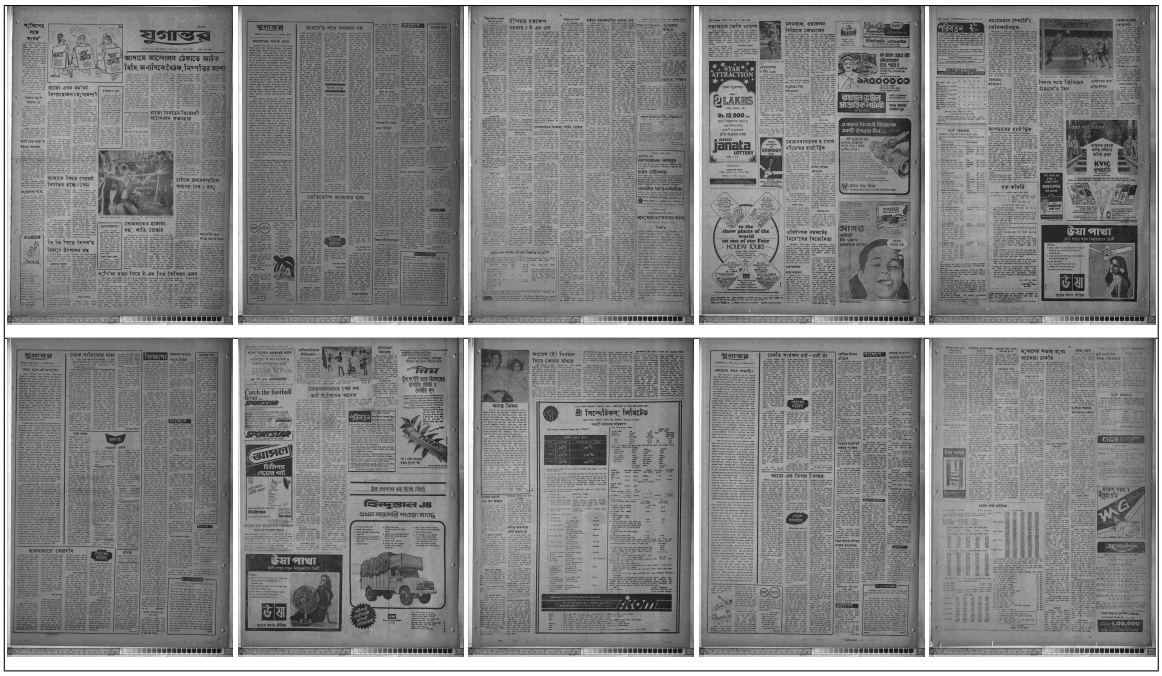


Table 8.1 : Sample Bangla newspaper dataset images

encoder and both segmentation and reconstruction branches. Newspaper images in our dataset are degraded noisy images and this poses a challenge for the model to learn robust features. Therefore, we train our model on clean Hindi newspaper images and then fine tune it for Bangla Newspapers. It facilitates teaching our model to comprehend layout syntax of the used dataset. This fine tuning gave us best result among other experimental settings. We term this model as M-proposed.

We compare our result with U-Net (Yin et al. [2022]) and Fully Convolution Network (FCN) proposed in (Meier et al. [2017]). OCR of noisy images are erroneous, therefore, we did not consider the OCR embedding from FCN architecture while comparing.

8.5 RESULTS AND DISCUSSION

Qualitative results for segmentation are shown in table 8.3 which shows results obtained by using two variants of our model viz. M-seg and M-seg-rec on 3 examples. The results show that our M-seg-rec model gives improved performance. The improvement over M-seg was primarily due to the addition of a reconstruction branch which enabled efficient feature extraction. It is worth noting that the dataset is quite difficult since the inter and intra article spacing are almost similar as can be seen in original images. Also, articles tend to merge at some points and this makes segmentation difficult.

We also report quantitative measures such as IoU (Intersection over Union) metric and accuracy. IoU is computed over ground truth label map and the output segmented region in the image. Table 8.4 gives the results reported in terms of these metrics.

8.5.1 Effect of denoising on segmentation

We demonstrate the effect of preprocessing the input images in our task of article segmentation. We hypothesize that the noise present in the heavily degraded Bangla newspaper



Table 8.2 : Sample Hindi newspaper dataset images

images will have an adverse effect on segmentation result. So we compare the segmentation results obtained for the noisy images and those obtained for denoised images. We applied denoising and super-resolution as pre-processing steps to the newspaper images before performing article segmentation. Table 8.5 pre-processed newspaper images obtained after denoising and super resolution.

Table 8.5 shows a comparison between segmentation results obtained for denoised and those obtained for noisy newspaper images. The improvement in segmentation performance is also evident from Table 8.6 which shows the IoU values and accuracy for segmentation of noisy and denoised images.

8.6 CONCLUSION

We test the effectiveness of the selective denoising and SinGAN-based super-resolution in improving the performance of the downstream task of segmenting noisy degraded Bangla newspaper images (taken from the National Digital Library of India) into constituent articles. An article constitutes its heading and body, and occasionally the pictures are embedded in the article. All other components such as comics, advertisements are excluded from the segmentation task and are considered as background. For better training of the newspaper segmentation model we also collect a dataset of clean newspaper images. We trained the model on both noisy and clean images and further fine tune it for noisy images. A comparison of the performance obtained by the trained model on the noisy and the denoised images showed a better segmentation performance for the denoised images. The nonuniform noise in the low resolution Bangla Newspaper images was effectively handled by the selective denoising technique.

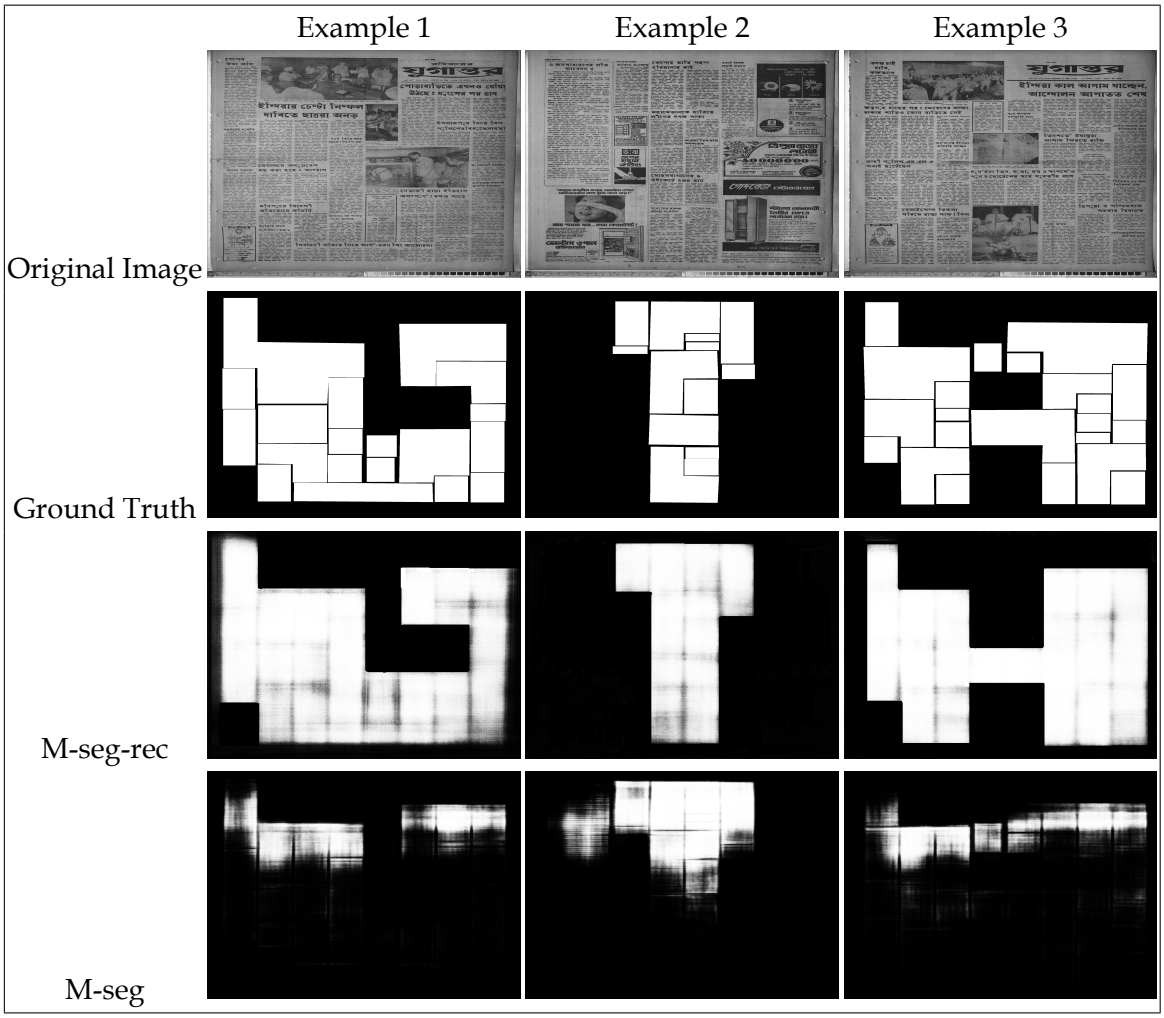


Table 8.3 : Segmentation results on Bangla Newspaper Images

Method	IoU	Accuracy
M-seg-rec	0.61	0.78
M-seg	0.48	0.65
FCN (Meier et al. [2017])	0.56	0.68
UNet (Yin et al. [2022])	0.57	0.76

Table 8.4 : IoU values obtained on Bangla Degraded Newspaper images.

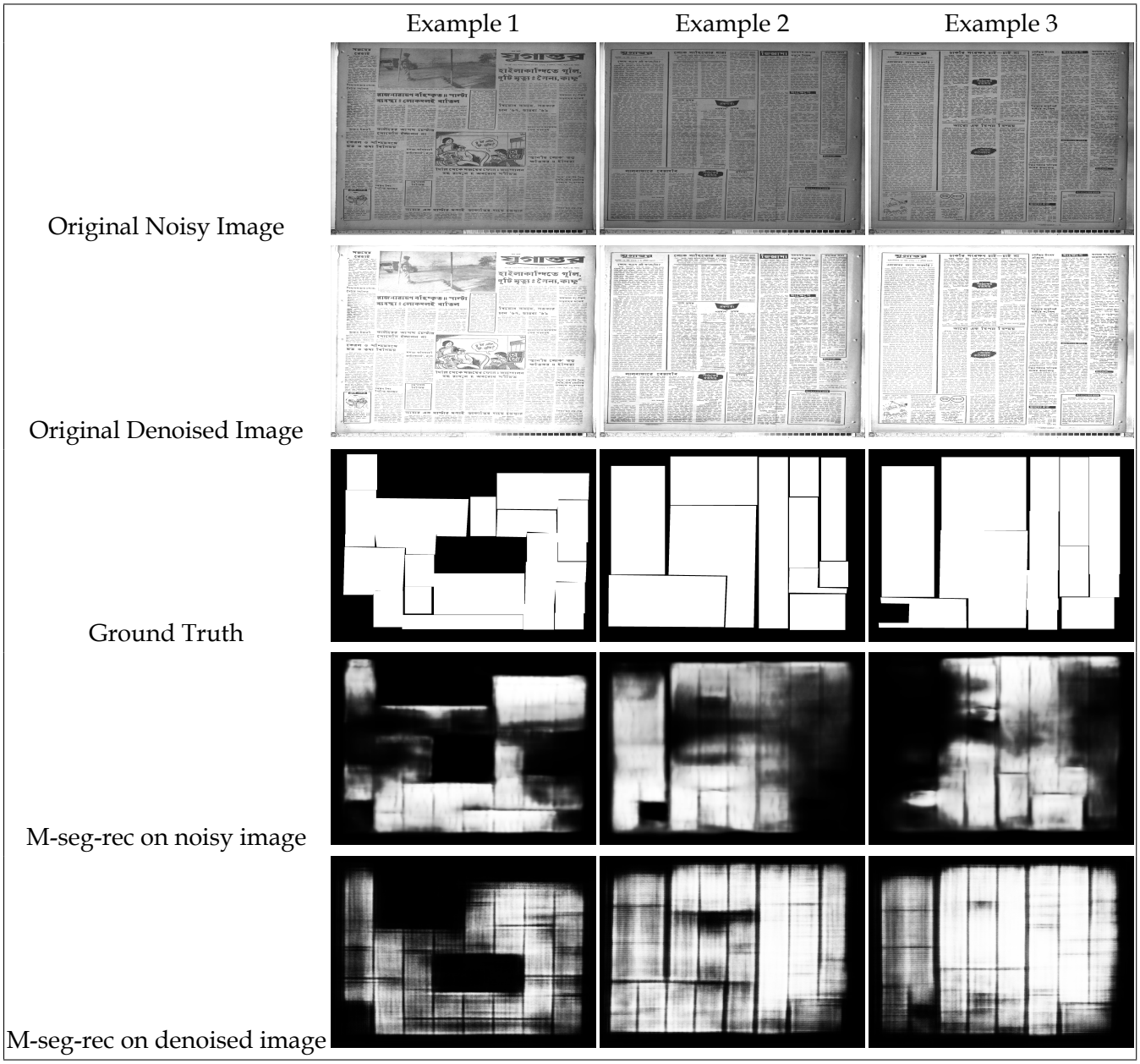


Table 8.5 : Segmentation results on Bangla Newspaper Noisy and Denoised Images

Method	IoU	Accuracy
M-seg-rec on noisy image	0.61	0.78
M-seg-rec on denoised image	0.73	0.85

Table 8.6 : IoU values obtained on Bangla Newspaper Noisy and Denoised images.