The primary purpose of this research is to implement Deeplabv3 architecture’s deep neural network in detecting and segmenting portable X-ray source model parts such as body, handle, and aperture in the same color scheme scenario. Similarly, the aperture is smaller with lower resolution making deep convolutional neural networks more difficult to segment. As the input feature map diminishes as the net progresses, information about the aperture or the object on a smaller scale may be lost. It recommends using Deeplabv3 architecture to overcome this issue, as it is successful for semantic segmentation. Based on the experiment conducted, the average precision of the body, handle, and aperture of the portable X-ray source model are 91.75%, 20.41%, and 6.25%, respectively. Moreover, it indicates that detecting the “body” part has the highest average precision. In contrast, the detection of the “aperture” part has the lowest average precision. Likewise, the study found that using Deeplabv3 deep neural network architecture, detection, and segmentation of the portable X-ray source model was successful but needed improvement to increase the overall mean AP of 39.47%.

Keywords: deep neural network, semantic segmentation, object detection, atrous convolution, local features

1. Introduction

In an unstructured or bomb-surrounded environment, manual operation of bomb detection and portable X-rays is a dangerous task, where a mobile X-rays system comprises an X-ray generator and an image capture unit or film [1, 2]. It processes the image in a remote location to determine whether an actual bomb exists inside random suitcases and containers. Nevertheless, to avoid the degradation of image quality from a portable X-ray system, one factor is to consider the alignment of the X-ray source to the film and maintain a distance within 12 mm or less between the target and film, as shown in Fig. 1, according to [2]. Likewise, detecting the X-ray source aperture for a sample image is shown in Fig. 2, and the distance of the film position from the suspect is the main challenge of alignment; thus, automatic alignment of the manipulator is necessary [1, 3, 4].

However, automatic alignment or manipulation requires reliable object detection and classification [5]. The camera’s aim may be lost, preventing the end-effector from performing the required task of detecting and aligning with the target item, such as the aperture of an X-ray model [1, 6]. Object detection usually recognizes a particular object within a scene, called global image features,
and the point of interest (POI) in the object is called local image features [7]. Global features can generalize a single vector over an entire object [8, 9]. In contrast, local features calculated at multiple points in an image are more robust to occlusions and clutter [8]. Therefore, the robustness of object detection decreases the probability that an object is lost for easy alignment. Similarly, a novel deformable part-based model takes advantage of the local context around each candidate detection and global context at the scene level [9]. Consequently, context-based approaches to perceptual inference tasks, such as object detection and semantic segmentation, accomplish perceptual inference tasks more quickly and accurately [9–13].

The primary goal of this study is to tackle the problem of determine the exact sections of the aperture, body, and handle of a portable X-ray source model with the scenario presented in Fig. 2.

Similarly, the aperture is smaller and has a low-resolution image, making it more difficult to segment with deep convolutional neural networks (DCCNs) [5, 9, 14]. Information regarding the aperture or object on a smaller scale can be lost as the input feature map shrinks as the net proceeds. To address this issue, it is recommended to employ atrous convolution, such as Deeplabv3, which is successful for semantic image segmentation [15–17]. As a result, in a portable X-ray model, a deep neural network with Deeplabv3 architecture segmentation is utilized to detect and locate the aperture, body, and handle.

Deep learning (DL) models, on the other hand, have delivered a new generation of image segmentation models with outstanding performance in recent years [11, 18]. Improvements frequently achieve the best accuracy rates on major benchmarks, leading to field paradigm changes. Image segmentation is a critical component of many visual understanding systems [12, 16, 19]. This entails dividing the images into several pieces or objects. Fig. 3 shows the picture segmentation outputs of a typical deep learning model using Deeplabv3 [12].

Image segmentation can be a pixel classification problem with semantic annotations or object partitioning. Semantic segmentation involves pixel-level labeling for all image pixels using a set of object categories, making it a more difficult task than image classification, which predicts a single label for the entire image [12, 16]. By detecting instances, instance segmentation broadens the scope of semantic segmentation and highlights each image’s POI.

2. Theoretical Basis

Object detection is the task of identifying cases within an object image. Determining whether an object exists in test images and locating the object is a fundamental problem in computer vision [14, 20, 21]. Object detection is important for many applications, such as robotic visual servoing, object tracking, video surveillance, security, mobile, medical application, satellite application, and internet services [5, 6, 22]. The algorithm must be sufficiently robust for detecting objects. To be more accurate, the orientation and scale of the descriptor and illumination state should be robust, and the object size and orientation should be changed during robotic manipulator motion. Problems can occur if the camera loses its object during servo operation [23]. Therefore, the robustness of object detection decreases the probability that an object will be lost. Object classification and localization are commonly used to recognize certain objects within a scene, recognizing the entire object called global image features or the POI in the object called local image feature [5, 8, 18]. Global features can be used to generalize a single vector over an entire object. In contrast, local features calculated at multiple points in an image are more robust to occlusions and clutter [8]. Most object recognition systems have adopted one of two approaches, exclusively using global or local features [8]. For instance, Cho and Kim proposed object detection using local features for robotic manipulator using adaptive binarization and differential [24]. The Gaussian method is applied to extract edge information from each image [10, 25]. It involves first extracting features such as contour lines from the object using the Fast Hough Transform [26] and selecting candidate lines to become the object’s geometric conditions. The specific region of the object is detected using the four vertices computed by the candidate lines using shape analysis [24]. Several edge extraction techniques, such as Sobel [27] colored image edge detection [28], Canny [29], and Gaussian methods [30–32]. However, the outline information of the object region cannot be extracted precisely because the lighting condition leads to distorted contrast and causes the object region to lose its features. Wang et al. introduced object recognition and pose estimation for the recognition engine of a modular manipulation system by using local features [20]. The goal of the recognition engine is to recognize the object placed in the task environment using sensor devices to provide other modules with information about the object for further object manipulation.
The local feature-based approach was used as a baseline method because of its simplicity of implementation and the need for a small training dataset [31]. Other common local features used corners as interest points from the digital image, where several edges intersect or an edge changes its direction, such as the SUSAN corner detector [32] and the Shi-Tomasi corner detector [33]. In contrast, the ridge descriptor approach computes a gray-level image, which can be viewed as a generalization of the medial axis. In addition, blobs use points or regions in the image that differ in particular properties such as brightness or color from the surrounding area, such as the principal curvature-based region (PCBR) detector and the maximum stability of extreme regions (MSER) [34]. Guo et al. further discussed existing 3D object recognition methods based on local surface features, including 3D keypoint detection, the description of local surface features, and surface matching [35]. Similarly, methods for global features have been widely used in 3D shape retrieval and classification [36]. Examples include geometric 3D moments [32], shape distribution [33], and characteristic viewpoint histograms [37, 38]. Therefore, they are not suitable for the recognition of a partially visible object in cluttered scenes. Sami et al. used shape-based algorithms to determine the object pose in the real position and orientation of the world coordinates of the moving target object. On the other hand, Lisin et al. introduced the possibility of combining local and global features for object class recognition using two methods as stacking ensemble technique and a hierarchical classification system [8].

In recent years, deep learning has become increasingly popular in computer vision, resulting in many object detection and semantic segmentation improvements [39]. When applied in a wholly convolutional form, DCNNs are helpful for semantic segmentation [23, 40]. Furthermore, because deep learning models have been successful in many vision applications, much research has recently focused on developing deep learning-based image segmentation algorithms [41]. For instance, a cascade of bottom-up images [42, 43], DCNN regions [23] convolutionally computed DCNN features for dense image labeling [44, 45] and category-level pixel labels [46].

The field of semantic segmentation has advanced dramatically since the first version was developed by Chen et al. made this public [16]. Semantic segmentation significantly increased the PASCAL VOC 2012 semantic segmentation benchmark [9]. For instance, scale-wave semantic segmentation [47], piecewise training of deep structured models [48], weakly and semi-supervised learning [49], boxsup [50], deep parsing network [49], and task-specific edge detection [50]. Surprisingly, most top-performing approaches have embraced the essential elements of the DeepLab system, called atrous convolution, for dense feature extraction and refinement of raw DCNN scores using a fully linked conditional random field (CRF) [15, 51, 52]. Atrous convolution, also known as dilated convolution, overcomes DCNN’s limitations of DCNN in applying semantic image segmentation, such as reduced feature resolution owing to pooling operations and invariance to local image transformation, which impedes dense prediction tasks requiring spatial information [17, 18]. Without requiring additional learning parameters, atrous convolution allows one to alter the resolution at which the feature responses are generated within the DCNNs, as shown in Fig. 4. Similarly, atrous convolution with input signal, $x(i)$ defined as

$$y(i) = \sum_{k=1}^{K} x(i+r \cdot k)w(k). \quad . \quad . \quad . \quad . \quad . \quad . \quad \quad (1)$$

Another parameter was introduced to the convolutional layers, called the dilation rate ($r$), which corresponds to the stride with a filter $w(k)$ of length $K$. In Fig. 4, the rate is determined by the dilation factor of the filter. The receptive field of the filter expands as the rate increases. Atrous convolution is equivalent to standard convolution at a rate of one.

Atrous or dilated convolutions have become prominent, and numerous recent publications have reported their real-time segmentation applications, such as DeepLab. Chen et al. developed DeepLabv1 [53] and DeepLabv2 [54], which are two of the most widely used image segmentation methods. The latter has three distinguishing characteristics. The first is dilated convolution to address the declining resolution of the network owing to max-pooling and striding. The second method is atrous spatial pyramid pooling (ASPP), which probes an incoming convolutional feature layer with filters at multiple sampling rates, capture objects and image context at various scales, and robustly segment objects at multiple scales [54]. The third improvement is improved object localization. Deep CNNs and probabilistic graphical mod-
els were used to create new limits. Chen et al. introduced Deeplabv3, which combines cascaded and parallel-dilated convolutional modules [17]. The ASPP contains all parallel convolution components. In ASPP, there is a one-to-one convolution and batch normalization. Concatenation of all outputs to construct the result with logits for each pixel and application of a $1 \times 1$ convolution [17, 18].

3. Method

The detection of the portable X-ray source model from the viewpoint of a semi-autonomous robot camera is the first process required for the successful alignment of the X-ray film. Deep neural networks were utilized consequently, there was a need to prepare a dataset for training. This section discusses the portable X-ray source model and then proceeds to dataset preparation, followed by the utilization of the deep neural network architecture and training.

3.1. The Portable X-Ray Source Model

The portable X-ray source model is a replica of the actual portable XR200 X-ray source developed by Golden Engineering [2] except for its functionality. Autodesk Fusion 360 [55] was used to model the portable X-ray source and convert it into 3D-printable model files. A comparison between actual and 3D-printed mobile X-ray sources is shown in Fig. 5.

A 3D-printed replica was made primarily because of the unavailability of an actual portable X-ray source, which is used for active service in police operations. Furthermore, this study also aims to investigate the flexibility of object detection using semantic segmentation algorithms by evaluating how the model can be generalized to the actual object of interest.

3.2. Dataset Preparation

The preparation of the dataset involves the following steps: (1) RGB and depth data capture of the portable X-ray source model, (2) raw data conversion and frame extraction, and (3) annotation and augmentation.

Intel Realsense D435i (Fig. 6), an inertial measurement unit, was used to capture the RGB and depth data. The software development kit (SDK) and documentation of the device are available at [57]. SDK provides an application that can directly interface with the RGBD camera and record it in an ROS bag file format. The application also allows frame-by-frame playback of the recording and 3D point cloud visualization.

The portable X-ray source model was placed under different lighting conditions in other scenes, whether partially occluded. The RGBD camera was then handheld while rotating around the mobile X-ray source model at most a meter away. There were at least forty separate recorded data captures of the portable X-ray source model, most of which resemble a 360° scan of the object, because acquiring as much data is necessary for the deep learning model. Individual frames, both RGB and depth, were extracted from each of the recordings, and at least 100. Each frame has a resolution of $1280 \times 720$ pixels and is saved as a separate image files resembling a timestamp. This was performed to match each RGB frame properly to its corresponding depth frame. Frames that did not contain occlusions were selected for segmentation mask annotation. Each specified frame was annotated by three different masks corresponding to the parts of the portable X-ray source model: (1) body, (2) aperture, and (3) handle. The object was partitioned for further applications; however, the main objective was to detect the aperture for X-ray film alignment. An example of an annotated frame is shown in Fig. 7. Here, the B mask overlay represents the body, H represents the handle, and A represents the aperture. All annotations were performed using the Hasty [58] (free account) cloud-based application.

Each RGB image and its corresponding annotated segmentation mask are structured into folders in PASCAL VOC format [17, 57]. All the images were resized to 50%...
Rogelio, J. P. et al.

Fig. 8. Comparison of without and with atrous convolution applied to cascaded layers [16].

of their original size to make the subsequent training process as short as possible.

3.3. Object Detection and Segmentation Model

This study was chosen to perform semantic segmentation using the Deeplabv3 architecture [15], which is suitable for detecting objects at various scales or distances. The Deeplabv3 architecture utilizes the atrous convolution method, which significantly improves the ability of the deep neural network to perform semantic segmentation at various object scales. The study also explored applying different atrous rates and parallelizing them. Using the PASCAL VOC 2012 validation set, Deeplabv3 attains a mean intersection-over-union (IoU) as high as 77.21%. An illustration of the comparison of without and with the atrous convolution architecture is shown in Fig. 8.

The Deeplabv3 architecture is readily available as a Tensor Flow model [54]. Because the supported dataset format output differs from that required for training the model, revising the program for training and evaluation is necessary. Of all the annotated frames, 80% were dedicated to training, and the rest were dedicated to validation.

4. Results and Discussion

4.1. The Portable X-Ray Source Dataset

The portable X-ray source dataset consisted of captured RGB data with the corresponding segmentation masks. The segmentation masks for each component were combined into a single image, as illustrated in Fig. 9.

Thus, proper mask layers must be observed for cases in which one component is in front of another. This also provides orientation information for the portable X-ray source from the viewpoint of the camera. The dataset also includes a JSON file format for each frame that describes the coordinates of the vertices of the mask and bounding box for each component. These vertices can be used to select the corresponding depth image pixels to further investigate how the depth can be used to improve segmentation. Likewise, bounding box information would be of great use for applications that only require the detection of the components or the whole object itself.

4.2. Visualization of Object Segmentation Results

The set of images below shows the visualization of the masks predicted by Deeplabv3. The images in Fig. 10 show the samples taken from the validation set. Qualitatively, the trained network could predict the masks for the “body” part. The masks for “handle” and “aperture” parts were also predicted for some cases. Note that the mask for the “body” is labeled as B, the mask for the “handle” is labeled H, and the mask for the “aperture” is labeled A. A quantitative analysis of the results is presented in the next section.
Fig. 10. Segmentation results from Deeplabv3. The left column shows the original RGB images, the right column shows the predicted segmentation masks.

Table 1. Mean IoU for predicted segmentation mask.

<table>
<thead>
<tr>
<th>Part</th>
<th>Percentage present</th>
<th>Percentage predicted present</th>
<th>Mean IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body</td>
<td>100.00%</td>
<td>98.91%</td>
<td>76.42%</td>
</tr>
<tr>
<td>Handle</td>
<td>100.00%</td>
<td>47.83%</td>
<td>10.52%</td>
</tr>
<tr>
<td>Aperture</td>
<td>92.39%</td>
<td>3.26%</td>
<td>0.79%</td>
</tr>
<tr>
<td>Overall mIoU</td>
<td>–</td>
<td>–</td>
<td>29.25%</td>
</tr>
</tbody>
</table>

4.3. Evaluation Metrics

4.3.1. Mean IoU for Segmentation

The mean IoU is the average of all pixel-wise IoUs for each part of the portable X-ray source. The pixel-wise intersection-over-union (IoU) metric is calculated by first obtaining the number of overlapping pixels (OLF) between two images and dividing it by the sum of the target pixels (TP) and predicted pixels (PP) from both images minus the number of overlapping pixels, as shown in Eq. (2). Subsequently, the overall mean IoU is obtained by averaging the mean IoU of each part.

\[
IoU = \frac{\sum OLP}{\sum TP + \sum PP - \sum OLP}.
\] (2)

Table 1 shows the mean IoU for each part, percentage of the validation set where the part is supposed to be present, and percentage of the validation set where the part is predicted to be present. The network achieved a higher prediction percentage for the “body” part of the portable X-ray source. On the other hand, the network has lower prediction accuracy for “handle” and “aperture” parts. In terms of the mean IoU, the network achieved a mean IoU that was very close to the highest possible that Deeplabv3 can obtain. The other parts have a significantly lower mean IoU, which can be attributed to their size and location relative to the larger “body” part areas.

4.3.2. Object Detection Recall, Precision, and Average Precision

Bounding boxes for each mask were obtained for the object detection metrics. A threshold value of 0.5 was used to determine if the detection yielded accurate positive results. Fig. 11 shows a side-by-side comparison of the predicted mask for each part on the left and corresponding bounding boxes on the right.

Figures 12, 13, and 14 show the precision-recall curves for detection of the portable X-ray source body, handle, and aperture respectively.
Table 2. Mean average precision for object detection.

<table>
<thead>
<tr>
<th>Part</th>
<th>Percentage present</th>
<th>Percentage predicted present</th>
<th>Mean IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body</td>
<td>100.00%</td>
<td>98.91%</td>
<td>76.42%</td>
</tr>
<tr>
<td>Handle</td>
<td>100.00%</td>
<td>47.83%</td>
<td>10.52%</td>
</tr>
<tr>
<td>Aperture</td>
<td>92.39%</td>
<td>3.26%</td>
<td>6.25%</td>
</tr>
<tr>
<td>mAP</td>
<td>–</td>
<td>–</td>
<td>39.47%</td>
</tr>
</tbody>
</table>

for the “body,” “handle,” and “aperture” parts, respectively, for the IoU threshold value of 0.5. Table 2 summarizes the average precision for each part, indicating that the detection of the “body” part has the highest average precision. By contrast, the detection of the “aperture” part has the lowest average accuracy. Overall, the mean AP was 39.47%, indicating that improvements had to be made.

5. Conclusion

The study showed that segmentation of the portable X-ray source was successful using DeepLabv3 as the deep neural network architecture. In this process, a mobile X-ray source dataset was prepared and will be of great use for further improvements in the study. Future directives include the utilization of the segmentation mask for depth images to further improve the accuracy of semantic segmentation and accommodate the possibility of occluded objects.

Acknowledgments

The authors would like to acknowledge the Department of Science and Technology, Metals Industry Research and Development Center, and the Intelligent Systems Laboratory of the De La Salle University for funding and helping us to finish this study.

References:


Name: Elmer P. Dadios  
Affiliation: Department of Manufacturing Engineering and Management, De La Salle University (DLSU)

Address: 2401 Taft Avenue, Malate, Manila 1004, Philippines  
Brief Biographical History: 1996 Received Ph.D. degree from Loughborough University 1997 Exchange Scientist, Japan Society for the Promotion of Science, Tokyo Institute of Technology 1998-1999 Director, Engineering Graduate School, DLSU 2000-2004 Director, School of Engineering, DLSU 2003-2004 General Chair of Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)  
Membership in Academic Societies: 1. Institute of Electrical and Electronics Engineers (IEEE) Computational Intelligence Society, Philippines, Founder and Chair 2. IEEE Region 10, Executive Member 3. The Mechatronics and Robotics Society of the Philippines (MRSP), Founder and President 4. IEEE, Senior Member 5. HNICEM, General Chair

Name: Ryan Rhay P. Vicerra  
Affiliation: Department of Manufacturing Engineering and Management, De La Salle University (DLSU)

Address: 2401 Taft Avenue, Malate, Manila 1004, Philippines  
Brief Biographical History: 2000 Received B.Sc. degree in Electronics and Communications Engineering from University of Santo Tomas 2001- Full-Time Faculty Member, University of Santo Tomas 2008 Received M.Sc. degree in Electronics and Communications Engineering from DLSU 2014 Received Ph.D. degree in Electronics and Engineering from DLSU  
Main Works: 1. Control systems, computational intelligence, and fuzzy logic.  
Membership in Academic Societies: 1. Institute of Electrical and Electronics Engineers (IEEE) 2. IEEE Computational Intelligence Society 3. IEEE Philippines Section 4. IEEE Computational Intelligence Society Philippines Chapter

Name: Argel A. Bandala  
Affiliation: Department of Electronics and Computer Engineering, De La Salle University (DLSU)

Address: 2401 Taft Avenue, Malate, Manila 1004, Philippines  
Brief Biographical History: 2008 Received B.Sc. degree in Electronics and Communications Engineering from Polytechnic University of the Philippines 2012 Received M.Sc. degree in Electronics and Communications Engineering from DLSU 2012- Full-Time Faculty Member, DLSU 2014 Received Ph.D. degree in Electronics and Engineering from DLSU  
Main Works: 1. Artificial intelligence, robotics, vision systems, swarm robotics, and multi agent systems.  
Membership in Academic Societies: 1. Institute of Electrical and Electronics Engineers (IEEE) 2. IEEE Computational Intelligence Society 3. IEEE Philippines Section 4. IEEE Computational Intelligence Society Philippines Chapter 5. IEEE Robotics and Automation Society

Name: Rogelio, J. P. et al.

Address: 2401 Taft Avenue, Malate, Manila 1004, Philippines  
Brief Biographical History: 1996 Received Ph.D. degree from Loughborough University 1997 Exchange Scientist, Japan Society for the Promotion of Science, Tokyo Institute of Technology 1998-1999 Director, Engineering Graduate School, DLSU 2000-2004 Director, School of Engineering, DLSU 2003-2004 General Chair of Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)  
Membership in Academic Societies: 1. Institute of Electrical and Electronics Engineers (IEEE) Computational Intelligence Society, Philippines, Founder and Chair 2. IEEE Region 10, Executive Member 3. The Mechatronics and Robotics Society of the Philippines (MRSP), Founder and President 4. IEEE, Senior Member 5. HNICEM, General Chair

Address: 2401 Taft Avenue, Malate, Manila 1004, Philippines  
Brief Biographical History: 2000 Received B.Sc. degree in Electronics and Communications Engineering from University of Santo Tomas 2001- Full-Time Faculty Member, University of Santo Tomas 2008 Received M.Sc. degree in Electronics and Communications Engineering from DLSU 2014 Received Ph.D. degree in Electronics and Engineering from DLSU  
Main Works: 1. Control systems, computational intelligence, and fuzzy logic.  
Membership in Academic Societies: 1. Institute of Electrical and Electronics Engineers (IEEE) 2. IEEE Computational Intelligence Society 3. IEEE Philippines Section 4. IEEE Computational Intelligence Society Philippines Chapter