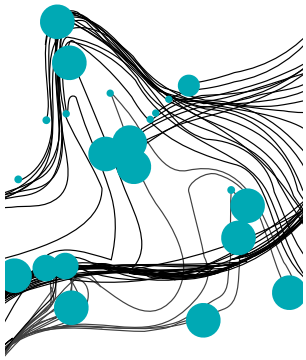
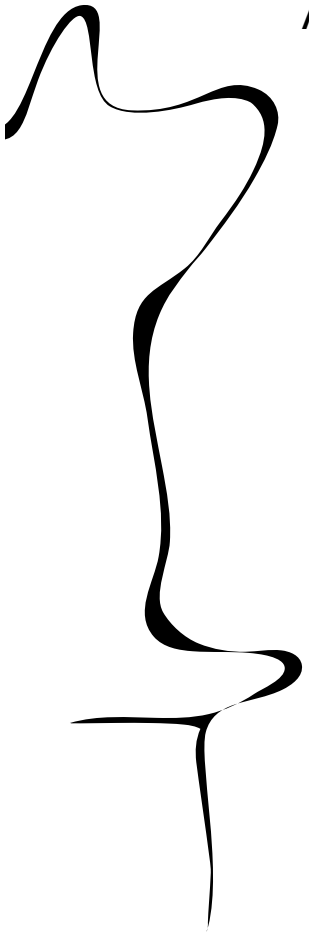


ROBOTICS AND MECHATRONICS



AUTOMATIC SEGMENTATION OF INDIVIDUAL SENSORS OF A MANOMETER INVIDEO-FLUOROSCOPIC VIDEOS FROM HEAD AND NECK CANCER PATIENTS

S.R. (Storm) Slurink

BSC ASSIGNMENT

Committee:
dr. F.J. Siepel
M.M. Rocha
A. Briassouli, Ph.D

June, 2025

043RaM2025
Robotics and Mechatronics
EEMCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | TECHMED CENTRE
UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Summary

Head and Neck cancer (HNC) patients often experience dysphagia (difficulty swallowing). To diagnose dysphagia it is important to evaluation the swallow quality. The golden standards for this evaluation are the videofluoroscopic swallowing study (VFSS) and fiberoptic endoscopic evaluation of swallowing (FEES). The problem with these kinds of studies are their reliance on subjective analysis of data, meaning that the accuracy of diagnosis greatly depends on clinician's expertise. High-resolution impedance manometry (HRIM) provides a more quantitive approach to analysing swallow quality. The altered anatomy of HNC patients can make it difficult to recognise the different anatomical regions needed to perform this analysis, however. To solve this issue, we can combine HRIM with VFSS to obtain additional information about the locations of these sensors and cross-reference these locations with the HRIM data. Cross-referencing the data from HRIM and VFSS manually is very time-consuming, however, so an automatic way of extracting the sensor locations from the VFSS video frame is necessary in order to streamline the process. This study attempts to develop such an algorithm using a ground truth mask containing the manometer as a starting point. We developed a technique based on adaptive thresholding, which was able to locate these sensors most of the time (F1-score = F1-score = 87.08 $\pm$ 15.55 for IoU $\geq$ 0.5). We also tried an approach using template matching but our method of extracting sensor locations from the template matching response map proved flawed. The adaptive thresholding algorithm was able to accurately determine sensor length (average error = 1.43 $\pm$ 1.13 mm) and sensor centres (average error = 1.08 $\pm$ 0.58 mm). Which suggests that the algorithm described in this study could prove very successful in the future.

Table of Contents

1 List of Abbreviations

Abbreviation	Definition
AT	Adaptive Thresholding
CLAHE	Contrast Limited Adaptive Histogram Equalisation
CT	Computer Tomography
FEES	Fiberoptic Endoscopic Evaluation of Swallowing
GT	Ground Truth
HNC	Head and Neck Cancer
HRIM	High-Resolution Impedance Manometry
IBP	Intrabolus Pressure
IBP-slope	Intrabolus Pressure-slope
IoU	Intersection over Union
IPP	Impedance at Peak Pressure
NI	Nadir Impedance
NKI	Netherlands Cancer Institute
OD	Oropharyngeal Dysphagia
PNI	Pressure at Nadir Impedance
PP	Peak Pressure
SRG	Seeded Region Growing
TM	Template Matching
TNI-PP	Time from Nadir Impedance to Peak Pressure
VFSS	Video-Fluoroscopic Swallowing Study

2 Introduction

2.1 Clinical Background

2.1.1 Head and Neck Cancer

Head and neck cancer (HNC) broadly refers to cancers found in the oral cavity, sinonasal cavity, pharynx and larynx (see Figure 1) [1]. It is one of the more common and deadly types of cancer being the sixth most common type of cancer and making up about 9.3 percent of all cancer deaths in 2024 [2]. Due to the inherent location of HNC, the tumour itself or its associated treatment can often lead to dysphagia (difficulty swallowing) [3]. The most common type of dysphagia affecting HNC patients is oropharyngeal dysphagia (OD) which refers to dysphagia during the oral and/or pharyngeal phase of the swallow [4]. OD can often result in dysphonia, pain, respiratory obstruction, malnutrition, and an overall decrease in the quality of life of the patient [5]. So it is very important to properly diagnose OD in order to treat these complications or prevent further ones.

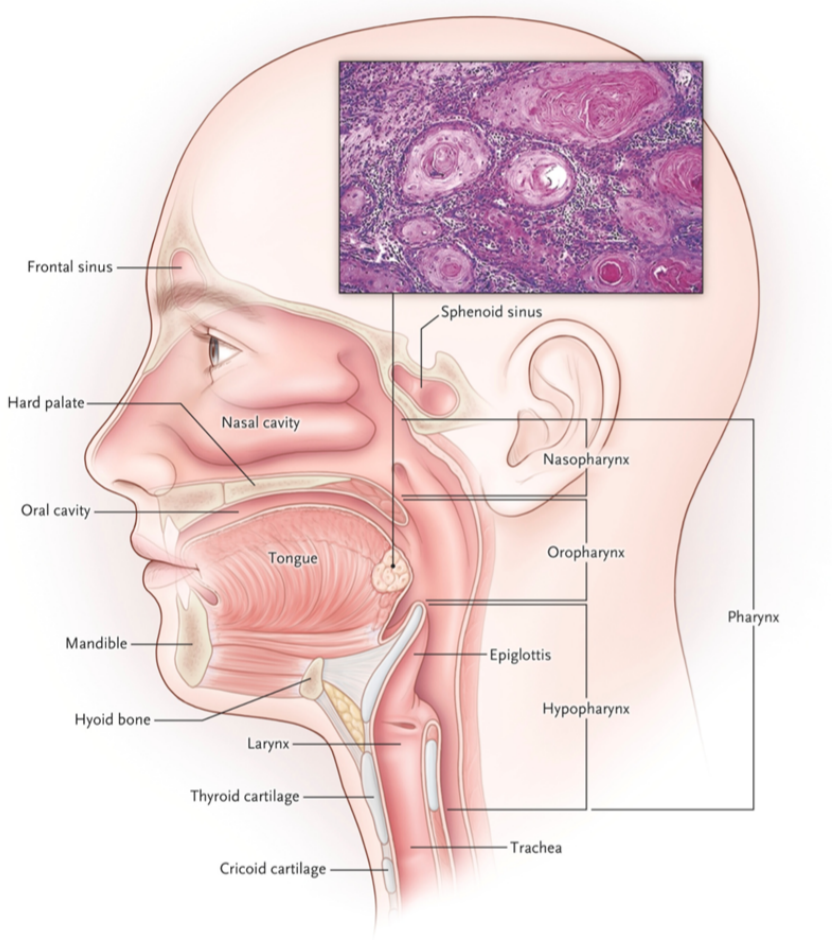


Figure 1: Schematic view of anatomical structures around the upper aerodigestive tract [1]

2.1.2 Current Diagnostics

The current gold standard in the determination of swallow quality and diagnosis of dysphagia are the videofluoroscopic swallowing study (VFSS) and the fiberoptic endoscopic evaluation of swallowing (FEES) [3]. VFSS is a real-time radiological study of the oral cavity, pharynx, larynx and oesophagus where the patient swallows a contrasting agent. This helps clinicians visualise the anatomical structures of the aerodigestive

tract and dynamics during the swallowing process (see Figure 2) [6, 7]. To diagnose OD, clinicians determine the swallow quality by looking for certain indicators [6]. Most importantly, VFSS allows for observing if parts of the bolus pass through the airways and to distinguish between penetration: the bolus enters the larynx but does not pass the vocal cords; and aspiration: the bolus passes through the vocal cords and enters the inferior airways. Other examples of indicators in the oral phase of the swallow include: labial competence; lingual control: the ability of the tongue to move the bolus towards the pharynx; palatoglossal closure: closing of the passage between the oral cavity and oropharynx by the palatoglossus muscle; presence of fractional deglution; patient needs to swallow multiple times in order to pass the bolus; and the presence of bolus residue in the oral cavity after the swallow [6]. Indicators for the pharyngeal phase include: inadequate palato-pharyngeal closure; triggering of the swallowing reflex when the bolus reaches the base of the tongue; hyoid and laryngial elevation; epilogical tilting; residue of the bolus in the pharyngeal cavity after the swallow; and abnormaities in the opening of the upper oesophageal sphincter [6]. VFSS allows for detailed examination of all these indicators, however this analysis can often be time-consuming [8]. Furthermore, the interpretation of these studies is subjective, meaning that accuracy of the diagnosis can be greatly influenced by the clinician's expertise, often leading to unreliable diagnoses [8].

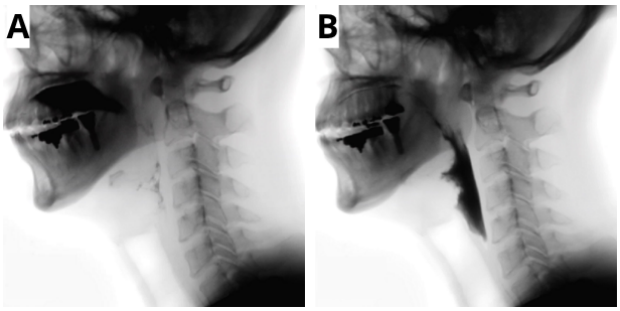


Figure 2: Two frames from a VFSS video from A) oral phase and B) pharyngeal phase [9]

FEES uses a flexible endoscope to asses the swallow from inside the pharynx, usually with the aid of coloured liquids or a solid bolus [10]. Due to the constriction of the pharynx, vision is obstructed during certain points of the swallow. Most notably, this makes it difficult to detect aspiration [11]. Much like VFSS interpretation of the data obtained from FEES is subjective, making resulting diagnoses inconsistent [12, 13].

2.1.3 High-Resolution Impedance Manometry

High-resolution impedance manometry (HRIM) offers great insight on the pharyngeal and oesophageal function and provides a more objective approach to characterising dysphagia [14]. HRIM uses a flexible catheter lined with numerous pressure sensors to measure the pressure and impedance along the pharynx and oesophageal sphincter during swallows [15]. The result of this measurement is a time-pressure plot where the y-axis indicates the sensor number, an example of which can be seen in Figure 3). From these plots, clinicians can derive a multitude of relevant variables including: Nadir impedance (NI): lowest impedance value, indicating the baseline impedance from the bolus; peak pressure (PP): the maximum recorded pressure; impedance at peak pressure (IPP): the bolus presence at maximum contraction; intrabolus pressure (IBP): pressure during luminal emptying; intrabolus pressure-slope (IBP-slope): rate of change of IBP; pressure at nadir impedance (PNI): IBP recorded when the bolus volume in the lumen is maximum; and time from nadir impedance to peak pressure (TNI-PP): the time from when the bolus volume in the lumen is maximum until peak contraction [16, 17]. These values need to be determined for the different anatomical regions involved in the swallow (see Figure 4) to increase the reliability of HRIM analysis [18].

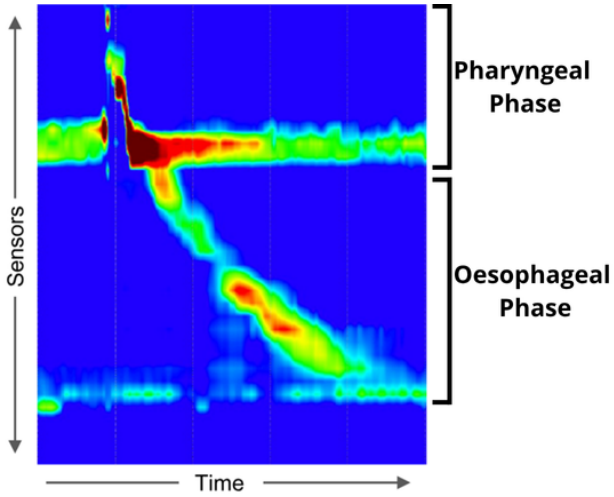


Figure 3: Example of a HRIM time-pressure plot of the pharyngeal and oesophageal phase of the swallow. The red areas indicate high pressure while blue areas indicate low pressure [19].

Although HRIM is able to provide more objective metrics for the diagnosis of OD, there are a couple limitations involved in this technique [18]. Firstly, there is currently no way to automatically perform this analysis. Secondly, the anatomical structures in the pharynx are not round, meaning the resulting pressure may not be accurate if the catheter is not in the middle. Finally, some data is lost when transitioning from a visual examination like VFSS and FEES to HRIM, for example it is not possible to detect aspiration or determine the amount of residue left after the swallow [18]. Another specific challenge with HNC patients is the fact that, due to the anatomical changes associated with the tumour or its treatments, the time-pressure plots become misshapen. The result is that it is no longer possible to distinguish the different anatomical regions needed to accurately determine the swallow quality. An example of this is provided in Figure 4, where plots of a healthy patient are compared to those of a HNC patient.

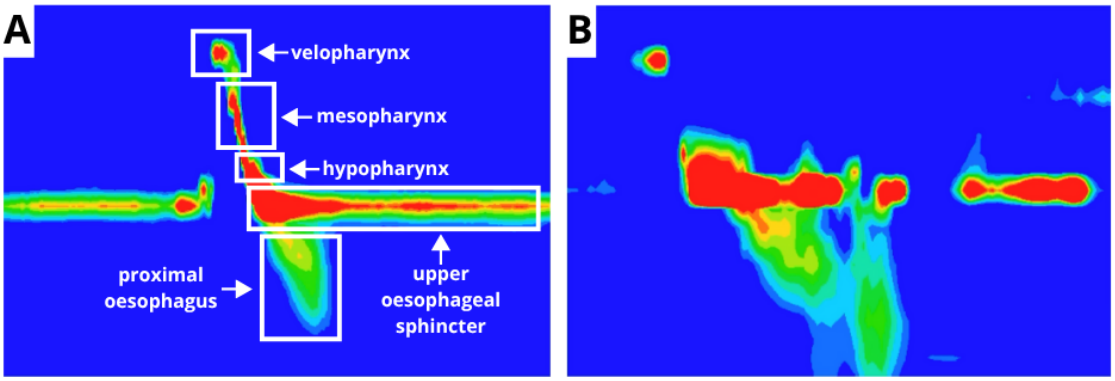


Figure 4: Comparison of two HRIM time-pressure plots of A) a healthy patient with distinct anatomic regions marked and B) a HNC patient with altered anatomy where distinct regions are not clear.

It should be possible to remove this last limitation by combining HRIM with VFSS in order to cross-reference the sensors in the time-pressure plots of HRIM and their anatomical location in the VFSS. Doing this manually is a time-consuming process however. Thus, an automated way to cross-reference these sensors could speed up this process significantly. Ultimately contributing to the automation of simultaneous HRIM and VFSS analysis, and a better OD diagnosis.

2.2 Technical Background

Image segmentation is the process of segmenting an image into different regions with the goal of making them more workable, processable and interpretable [20]. An example of this is given in Figure 5. In the world of medical imaging, image segmentation allows for more accurate diagnoses and pre-surgical planning [21].

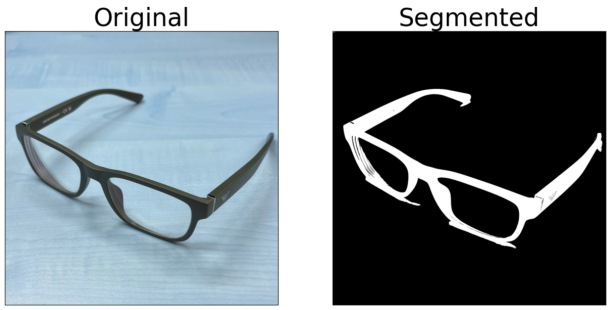


Figure 5: A simple representation of image segmentation using thresholding (see Section 2.2.1)

Although segmentation approaches based on deep-learning play an important role in the segmentation of medical images, this study will focus on traditional segmentation approaches due to the relative ease of their implementation and lower resource consumption [21].

Xu et al. list five main segmentation techniques in traditional segmentation of images: thresholding, edge-based segmentation, region-based segmentation, clustering-based segmentation and graphic-based segmentation [21]. These techniques along with template matching could be valuable tools in achieving proper segmentation of the manometer sensors.

2.2.1 Thresholding

Thresholding is a fairly basic technique where pixels of a greyscale image are individually evaluated and set to either 0 or 1 depending on if the intensity reaches a certain threshold [21]. Thresholding methods can be divided in two categories: global thresholding and local thresholding, though hybrid methods do exist. Global thresholding uses a single unchanging threshold for the whole image while local thresholding employs a variable threshold dependent on the characteristics of surrounding pixels [22]. Local thresholding, also referred to as adaptive thresholding, offers more control which can increase the segmentation quality in similar medical images [23].

Adaptive thresholding computes the threshold based on the mean of a specified window size around each pixel [24]. There are different ways to calculate these means, one such way is the use of Gaussian weights, where pixels nearer to the centre contribute more than those farther away. This threshold is mathematically described in equations 1 and 2. Here, $b(x, y)$ is represents the pixel at position (x, y) in the thresholded image, $T(x, y)$ represents the threshold value for the pixel at position (x, y), W is the window in which the mean is calculated, $G(i, j)$ is the Gaussian weight assigned to the pixel at position (i, j) in W, $I(i, j)$ is the greyscale intensity of the pixel at position (i, j) in W, and C is a constant used to manually adjust the threshold [24].

$$b(x,y) = \begin{cases} 0, & \text{if } f(x,y) \leq T(x,y) \\ 1, & \text{if } f(x,y) > T(x,y) \end{cases} \tag{1}$$

$$T(x,y) = \sum_{(i,j)\in \mathrm{W}} G(i,j) \cdot I(i,j) - C \tag{2}$$

2.2.2 Template Matching

Template matching involves comparing parts of the total image with a smaller, predetermined template. A response map is generated based on which parts of the image are most similar to the template. Different template matching algorithms often differ in how they measure the similarity between the image and the template. For instance, some of the simpler comparison methods calculate the total sum of absolute differences or cross-correlation coefficients [25].

2.2.3 Pre-Processing Techniques

Before applying any segmentation techniques, we can first apply different methods of pre-processing to enhance segmentation. A Gaussian blur can be applied to reduce image noise, for example [26]. Another useful technique is contrast limited adaptive histogram equalisation (CLAHE) [27]. CLAHE can be used to enhance contrast in an image. It does this by first dividing the image into multiple 'tiles' and then redistributing the intensity values within the image to span a wider range of intensities, usually between 0 and 255 [28].

2.3 Related Work

In the world of medicine, image segmentation and object detection are powerful tools for the identification of certain structures to assist in diagnoses. The aforementioned Xu et al. [21] outline many such examples, including ones using deep learning. Kiran et al. [23] also outline various segmentation techniques, including thresholding, clustering, and edge-detection methods, and compare their ability to segment the lungs from chest X-rays. Larhmam et al. [29] use Canny edge-detection and a generalised Hough transform to determine the locations of vertebrae in the spine from X-ray images.

For HRIM and VFSS specifically, Geiger et al. [19] have worked on HRIM catheter segmentation and sensor localisation in VFSS videos for similar purposes as outlined here. Their approach uses template matching (see section 2.2.2) to determine the highest probability sensor locations with great results. Their research is focused on sensor localisation during the oesophageal phase of the swallow, however, while this study focuses solely on localisation during the pharyngeal phase. Furthermore, their particular dataset has consistent image dimensions and as a result a consistent sensor size between images. We want a way to automatically determine the sensor length so sensor localisation can be done for a more wide variety of conditions.

2.4 Research Goal

Despite the fact that HRIM can provide a much needed quantitive approach to OD diagnosis, there are still limitations that prevent its use in the clinical practice for HNC patients. Firstly, there is the fact that there is currently no way to automatically analyse HRIM data [18]. Secondly, the anatomical changes caused by the tumour or its treatment can make it difficult to distinguish the multiple anatomical landmarks during the swallow, which is a crucial step in determining swallow quality using HRIM [18].

In this study we aim to develop an algorithm that is capable of extracting the HRIM sensor locations from VFSS video frames in order to cross-reference them to the HRIM data with these locations to make determining the anatomical landmarks easy and fast. An example of this is provided in Figure 6. In this study, we will extract the sensor locations from a mask of the HRIM catheter and not the entire VFSS image. Furthermore, this study will be solely focused on traditional segmentation techniques.

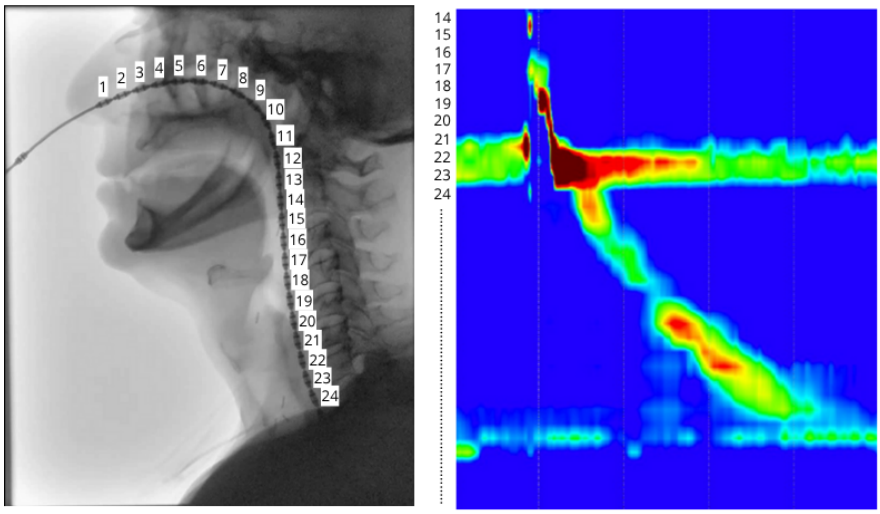


Figure 6: The sensor locations are detected and labelled in VFSS, then overlaid on the HRIM time-pressure plots so anatomical landmarks can be easily determined. (The VFSS frame and HRIM time-pressure plot in this figure are not from the same research or patient so the anatomy/pressure may not correspond perfectly and are merely used to provide an example of the research goal.)

3 Methods

3.1 Dataset

A total of 102 swallowing videos from 16 patients were provided from the Netherlands Cancer Institute (NKI). Two frames from each video were randomly selected resulting in 204 video frames. Along with these, the ground truth mask outlining the catheter was provided (cf. Figure 7) To further test the algorithm in real-world conditions we obtained the the catheter mask predictions from a different algorithm developed by Rocha et al. [30].

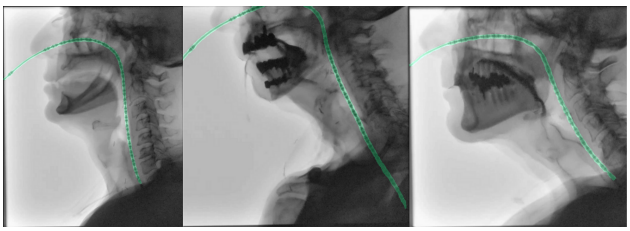


Figure 7: Example of numerous VFSS frames with the provided ground truth masks overlayed in blue.

3.2 Segmentation Algorithm

The segmentation can be roughly divided into four main steps: pre-processing, adaptive thresholding, template selection, and template matching. A general overview of the steps is given in Figure 8. The rest of this section will be dedicated to explaining each step in more detail.

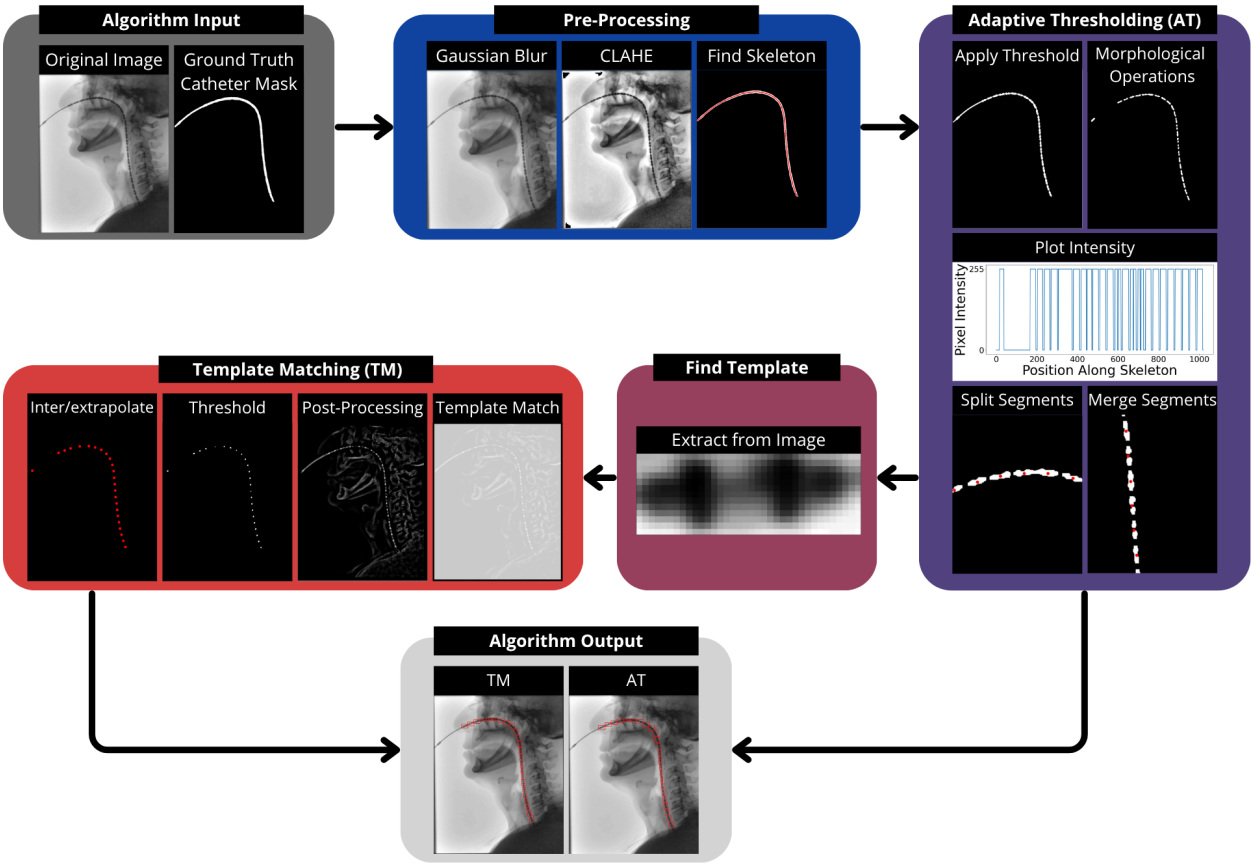


Figure 8: A general overview of the segmentation algorithm and its four main steps: pre-processing, adaptive thresholding, template selection, and template matching.

3.2.1 Pre-Processing

This algorithm assumes that the catheter region has been previously segmented. For testing purposes, the dataset used for this study contained the original VFSS video frames and their corresponding catheter annotation masks (see Figure 9). The first step is to prepare our images for processing. Firstly, we apply Gaussian blur to the image to reduce noise. Then we use Contrast Limited Adaptive Histogram Equalization (CLAHE) [27] to improve contrast within the image, with the goal to improve sensor visibility (see Figure 10). Applying CLAHE can cause undesired dark patches to appear in the background, so we remove these patches by setting the intensity of all pixels below a certain intensity value to 255. We do this because the sensors are darker than the background and we do not want these dark patches to interfere with the next step: adaptive thresholding. In this step we also obtain the skeleton of the catheter from the mask, that is to say a list of points describing a line that runs through the middle of the catheter. We use this skeleton in the next steps.

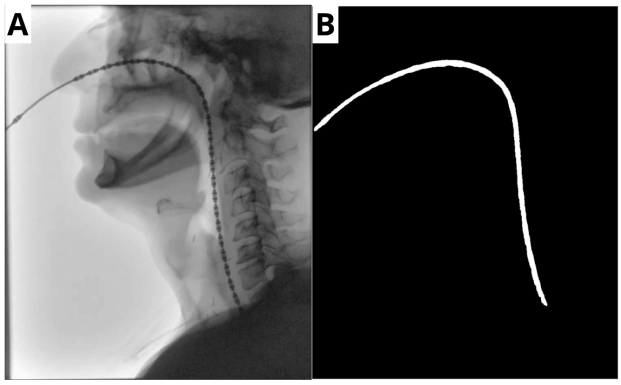


Figure 9: Startpoint of the algorithm: A) Original frame from VFSS video B) (Ground truth) mask containing HRIM catheter.

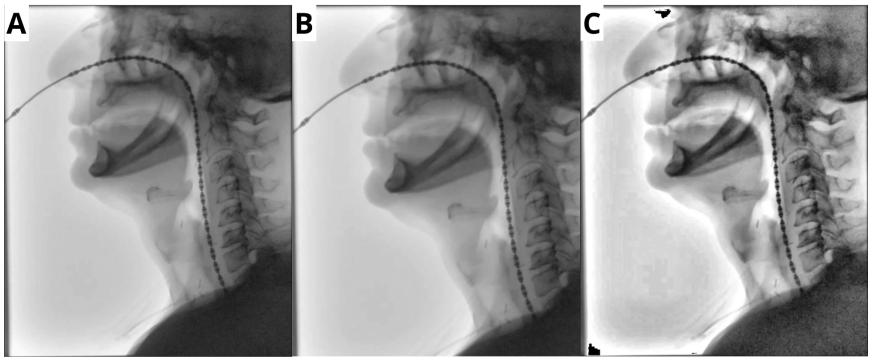


Figure 10: Pre-Processing used for sensor localisation. A) Original VFSS frame. B) Gaussian blur. C) CLAHE (cliplimit=2).

3.2.2 Adaptive Thresholding

An adaptive thresholding methodology is applied to the contrast-enhanced image. After this, we also limit the search region by applying the ground truth (GT) mask (cf. Figure 9b). An example of the result is shown in Figure 11b. We then perform numerous morphological operations to fine-tune the thresholding. Initially, we use a single morphological closing operation, where all white regions grow in each direction by one pixel, to get rid of any gaps. Then we perform numerous morphological opening operations, where all white regions shrink by one pixel. Finally, we apply a final closing operation to isolate each individual sensor. The result can be seen in Figure 11c. The number of morphological openings can differ between images. The base value is 4 but if the number of sensors after these operations is too low, which can happen if the opening operations get rid of sensors entirely, we reduce this number to 3. Conversely, when the number of sensors detected is too high, this likely means that the current number of morphological opening operations do not remove the part of the catheter where no sensors are present. To get around this we increase the number of opening operations to 5. The final closing operations are always performed two times. We also use thresholded image (cf. Figures 11b and 12a) to compute the orthogonal distance from each skeleton point to the background. We then use a Savgol filter to reduce noise. The result is a graph like shown in Figure 12b. Each sensor is an oval shape with two dark bars inside. In the thresholded image this means that each sensor is connected at a very thin point, and has a small indent in the middle of where the space between the two dark bars would be. In the plot of these distances (cf. Figure 12b) this translates to an "M"-like shape for each sensor. By detecting each of the lower valleys in the graph, we can estimate the sensor length l_{sens} by taking the median of the distances between these valleys.

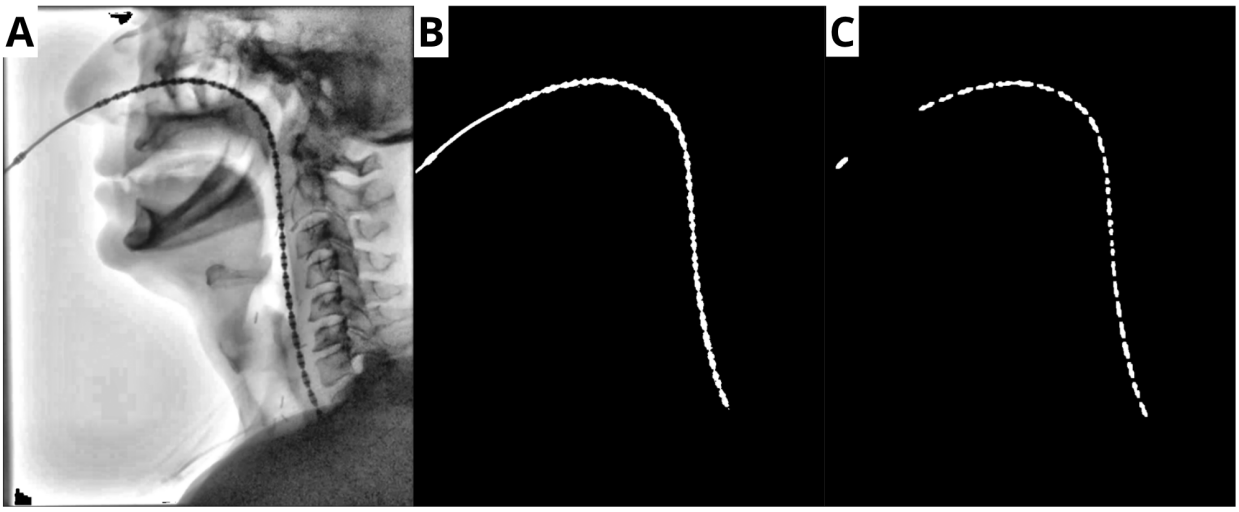


Figure 11: An example of a frame where adaptive thresholding is applied: A) Contrast-enhanced image obtained through CLAHE. B) Adaptive threshold with windowsize = 31, c = 2 (see Section 2.2.1) after limiting the search region with the ground truth mask (cf. Figure 9b). C) Adaptive thresholding after morphological operations: 1x closing, 3x opening and 2x closing.

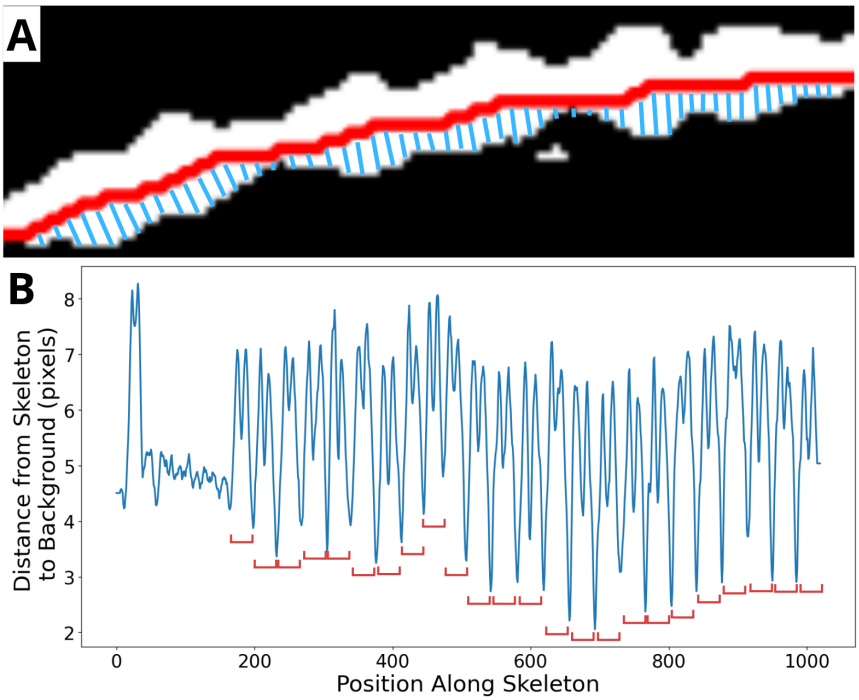


Figure 12: Orthogonal distance from the skeleton to the background: A) Numerous sensors where the orthogonal distance from the skeleton (red) to the background is visualised with blue lines. B) All distances plotted against their corresponding position along the skeleton, filtered with a Savgol filter. Each sensor is indicated by a red bracket.

We can now simplify the problem of locating each sensor by plotting the intensity of the thresholded image after performing the morphological operations (cf. Figure 14) for each point in the skeleton (cf. Figure 13). From this plot we can easily infer the start- and endpoints of each segment. So now we only need to

consider the intensity values for the points in the skeleton, effectively reducing the problem to one dimension. Each segment does not directly translate to one sensor, however. Sometimes these morphological operations can cause sensors to split in two or merge together (cf. Figure 11c) in the thresholded image, so we use l_{sens} to determine when to group these segments together or split them apart. We also determine when we detect only half a sensor, by looking if we have a single shorter segment next to a larger gap. We can also use this plot to further refine the number of morphological closing operations by determining the length of each detected segment, and inferring if the segmented image contains more half sensors than whole ones (determined by the estimated sensor length). In that case, we apply an additional morphological closing operation.

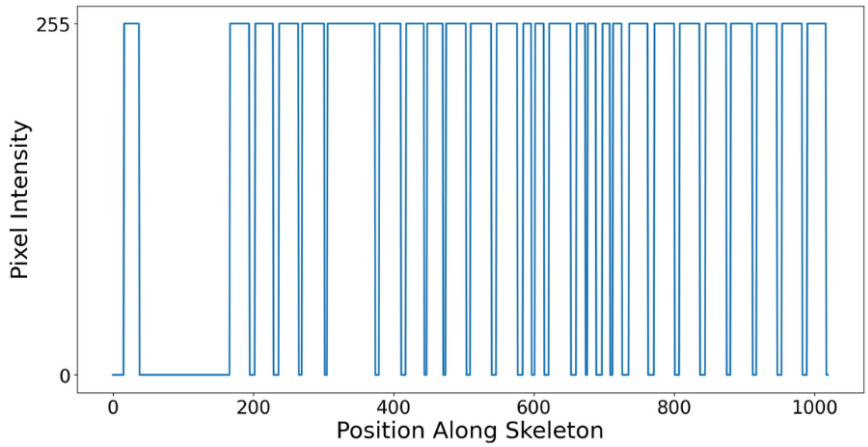


Figure 13: Plot of pixel intensity of the thresholded image after morphological operations (cf. Figure 2c) for each point in the skeleton.

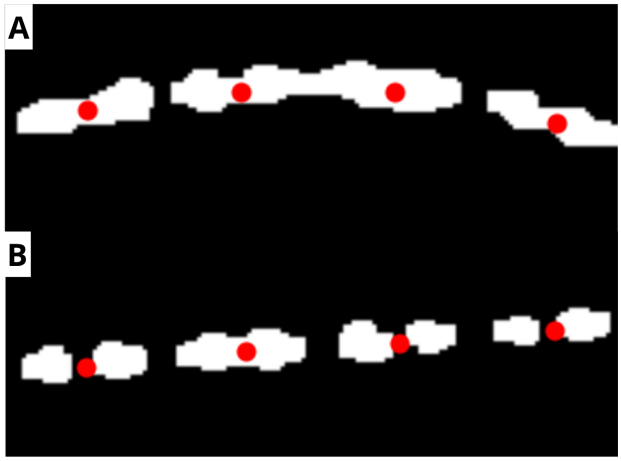


Figure 14: An example of splitting and merging of thresholded segments is necessary. The red dots denote the centres of the detected sensors. A) An example of a merged segment where we need to apply splitting. B) An example of multiple split segments where we need to apply merging.

As a final check we plot the intensity of the contrast-enhanced (CLAHE) image along the skeleton within each individual detected sensor (cf. Figure 15). Each sensor contains within it a dark-light-dark pattern (cf. Figure 15a). We want to make sure that the sensor we detected is actually a sensor. To do this we take the main peak and valleys from this intensity plot (cf. Figure 15b) and calculate the difference in intensity between the peak and the average of the valleys. When this difference is large enough we know this is a valid sensor.

This check mainly exists to remove the false positives on the part of the catheter where no sensors are present (cf. Figure 9a). This is only the case in before the first sensor on the catheter, so we only need to perform this check until we detect two valid sensors. We need to detect two to account for the reference sensor (cf. Figure 9a).

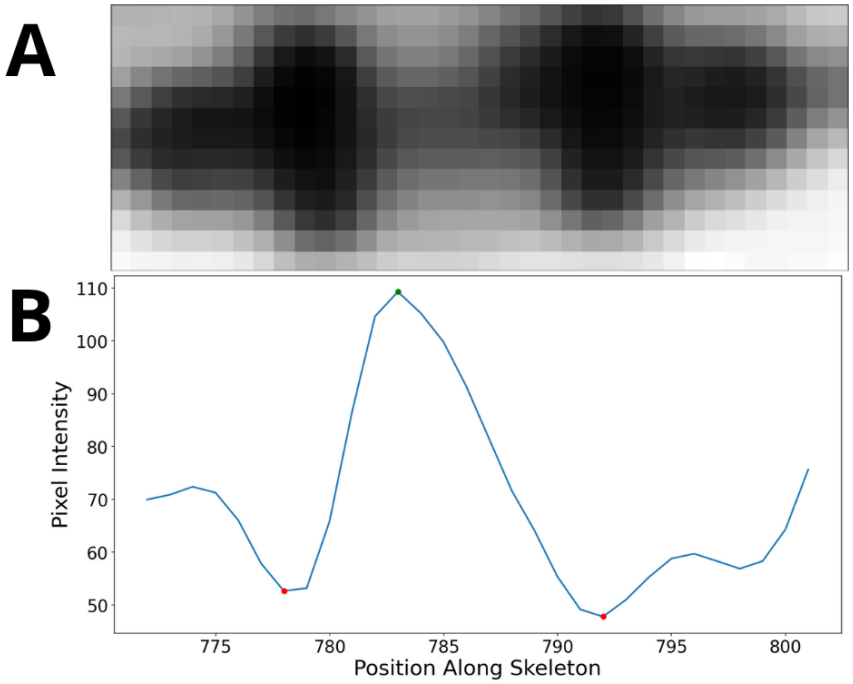


Figure 15: A) Example of HRIM sensor after applying CLAHE, clearly displaying the dark-light-dark pattern. B) Intensity of contrast-enhanced image along a part of the skeleton corresponding with a single sensor. The main peak is denoted by the green dot and the main valleys are denoted by the red dots.

3.2.3 Template Selection

In order to perform template matching, we first need to obtain the templates. As opposed to the work of Geiger et al. [19], we propose an automated template selection algorithm. For this, we refer back to the graph in Figure 15. We can use the sensor intensity plot to quantify the sensor quality, based on four parameters:

1. The difference in intensity between the peak and the average of the valleys I_{peak}; this should be maximized.
2. The difference in intensity between the two main valleys $I_{valleys}$; this should be minimized.
3. The difference in the distance between the first valley to the peak and the distance between the second valley to the peak $x_{valleys}$; this should be minimized.
4. The distance from the main peak to the centre of the plot x_{peak}; this should be minimized.

We use these values to assign each possible template a score and take the three highest-scoring templates. Equation 3 shows how these scores are calculated. w represents the weight of these parameters. We use: $w_1 = w_2 = w_3 = 1$ and $w_4 = 2$. The three best templates of the example frame are shown in Figure 16.

$$\text{Score} = w_1 \cdot I_{peak} - w_2 \cdot I_{valleys} - w_3 \cdot x_{peak} - w_4 \cdot x_{valleys} \tag{3}$$

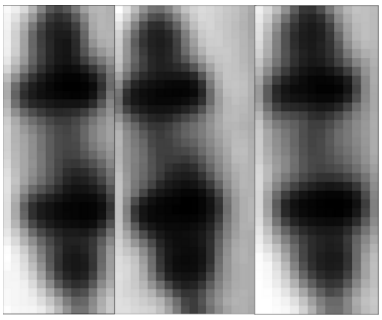


Figure 16: Templates extracted from the contrast enhanced-image (CLAHE) using the described algorithm.

The templates are then pre-processed to make sure they can be used for template matching. Since we will use python OpenCV's `matchTemplate` function, which does not have the ability to perform rotation-invariant template matching, we need to rotate the templates on our own. We obtain rotated forms of each template for a range between 0° and 180° with increments of 5°. We then pad each template for all rotations so all templates are the same size. This ensures the output of template matching is also always the same size. We also create a mask for each rotation so that we only use the non-padded parts of the image for template matching.

3.2.4 Template Matching

We use the templates obtained in the previous step and perform template matching on the contrast-enhanced (CLAHE) image by comparing the cross-correlation coefficient across all templates and rotations. For each pixel, we take the highest intensity value of all rotations and then take the average of all templates, resulting in a response map (cf. Figure 17a). To improve contrast in the response map, we remove all pixels below a certain intensity threshold and normalise the remaining values(cf. Figure 17b). Finally, we threshold the image to get the highest cross-correlation values along the search region defined by the GT mask (cf. Figure 9b). The result is shown in Figure 17c.

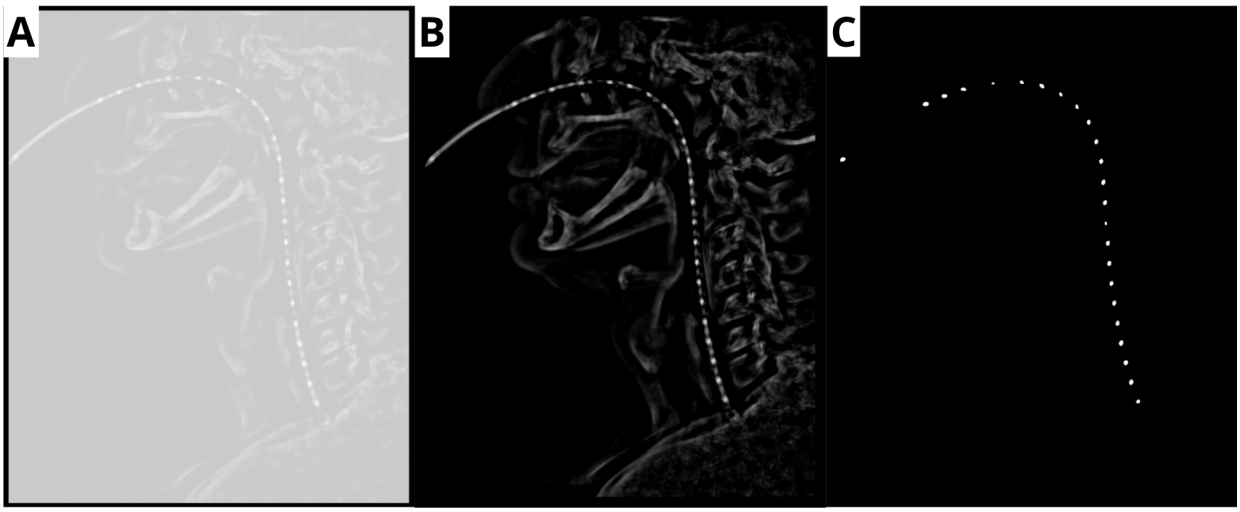


Figure 17: An example of a frame where template matching is applied: A) The template matching response map in which each pixel represents the maximum value across all rotations and the average across all templates. B) The response map after normalising the values. C) The response map after limiting search area with ground truth mask (cf. Figure 9b) and applying a threshold

Then, to once again simplify the problem of sensor localisation, we turn the problem one-dimensional by

taking the centres of all segments and compute the closest point within the skeleton. Next, we check the validity of each sensor. We do this by making predictions of where we think other sensors will be based on the average distance between points. We make three predictions on both sides of each segment and calculate if another sensor is present within a certain distance of the predictions. We give each segment a score based on how many predictions are within a certain distance of another detected segment in the response map. If two or more predictions are within acceptable range of another segment in the response map, we assume they are correct. If, after this, a two segments are still within a half l_{sens} distance to each other, we remove the segment with with fewer accurate predictions.

After removing all false detections, we can start inter/extrapolating the remaining sensors. We calculate the number of sensors that need to be inter/extrapolated based on l_{sens}. For extrapolation, the first and last sensors detected using adaptive thresholding are used to set the boundary regions where extrapolation is needed.

3.3 Validation

The bounding boxes were manually annotated for all 204 frames from the VFSS videos. These bounding boxes were compared to those given by the algorithm output using the intersection over union (IoU) metric (see Figure 18). When the IoU exceeds a predetermined threshold, the detection is considered correct. Using this we can calculate a number of parameters: the precision, recall and F1-score (Equations 4-6). Precision represents the amount of retrieved items that are correct, while recall (or sensitivity) represents the amount of correctly detected items. The F1-score is the harmonic mean between precision and recall.

We also calculate: the average distance between the centres of the true positive bounding boxes and ground truth bounding boxes for IoU ≥ 0.5, the error in the estimated sensor length l_{sens}, and the accuracy of the number of sensors detected without considering IoU (see Equation 7). The segmented frames are also manually inspected to ensure the results are accurate and representative.

We compute these parameters for four different algorithms:

- A baseline algorithm using only adaptive thresholding (only the steps described in Section 3.2.1 and 3.2.2) that also uses a manual length estimation method instead of the one described in Section 3.2.2, where the estimation is made by multiplying the image dimensions with a manually determined reference length.
- An algorithm using only adaptive thresholding with the ground truth mask.
- An algorithm that also uses the template matching steps described in sections 3.2.3 and 3.2.4.
- An algorithm that uses adaptive thresholding with the mask predictions instead of the ground truth masks.

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{4}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{5}$$

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

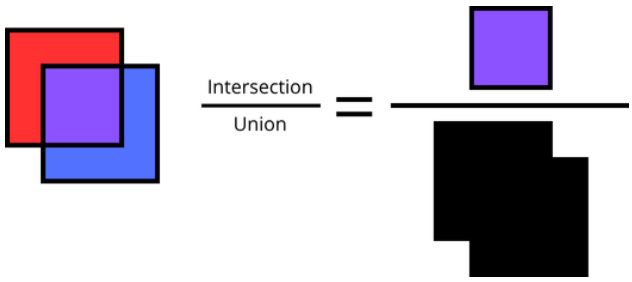


Figure 18: Intersection over Union of two bounding boxes.

$$\text{Accuracy} = \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}} \tag{7}$$

4 Results

For the results we make distinction between four different methods in the same way as described in Section 3.3. Again, they are:

- A baseline algorithm using only adaptive thresholding (only the steps described in Section 3.2.1 and 3.2.2) that also uses a manual length estimation method instead of the one described in Section 3.2.2, where the estimation is made by multiplying the image dimensions with a manually determined reference length. (AT-est).
- An algorithm using only adaptive thresholding with the ground truth mask (AT).
- An algorithm that also uses the template matching steps described in sections 3.2.3 and 3.2.4 (TM).
- An algorithm that uses adaptive thresholding with the mask predictions instead of the ground truth masks (AT-pred).

In Table 1 we see that the sensor length detection algorithm is able to estimate the rough size of the sensors, with an average error of 5.75 ± 4.79 pixels or 1.43 ± 1.13 mm when using the ground truth mask. The results are very similar when using the mask predictions with an average error of 1.43 ± 1.12 mm. We also see that the error in the centres of the correctly identified sensors (IoU ≥ 0.5) are fairly small, with an average of 1.08 ± 0.58 mm for the ground truth masks and 1.39 ± 0.73 mm for the predicted masks. The template matching algorithm performs worse in this aspect with an average error of 3.70 ± 2.35 mm. The same is true for the sensor count accuracy where AT and AT-pred perform relatively equally with an accuracy of 94.02 ± 7.40% and 92.51 ± 9.48% respectively but TM only has an accuracy of 83.90 ± 19.41%.

Table 1: The error in the estimated sensor length, accuracy of the number of detected sensors, and the average distance between the centres of the true positive detected sensors (IoU ≥ 0.5) and ground truth.

Method	Sensor Length Error (pixels)	Sensor Length Error (mm)	Sensor Count Accuracy (%)	Sensor Centre Error (mm)
AT	5.75 ± 4.79	1.43 ± 1.13	94.02 ± 7.40	1.08 ± 0.58
TM	”	”	83.90 ± 19.41	3.70 ± 2.35
AT-pred	5.71 ± 4.83	1.43 ± 1.12	92.51 ± 9.48	1.39 ± 0.73

Table 2 shows the precision, recall, and F1-score for different IoU thresholds for each method. The performance of the adaptive thresholding is very similar to the algorithm using manual length estimation with an F1-score of 88.21 ± 13.02% and 88.20 ± 12.32% respectively when IoU ≥ 0.5. The algorithm using the mask predictions instead of the ground truth masks has a slightly lower F1-score of 86.18 ± 14.48 when IoU ≥ 0.5. The template matching performs much lower than all other algorithms with an F1-score of 65.35 ± 37.11 for IoU ≥ 0.5. This pattern persists trough all IoU thresholds.

Table 2: Precision, recall and F1-score for different IoU thresholds organised per method.

Method	IoU Threshold	Precision (%)	Recall (%)	F1-Score (%)
AT Manual Length Estimation (AT-est)	0.50	88.79 ± 11.68	87.95 ± 13.65	88.20 ± 12.32
	0.55	84.94 ± 13.65	84.14 ± 15.18	84.38 ± 14.13
	0.60	77.93 ± 15.57	77.22 ± 16.74	77.43 ± 15.94
	0.65	66.28 ± 17.97	65.73 ± 18.79	65.88 ± 18.21
	0.70	53.81 ± 19.20	53.44 ± 19.80	53.52 ± 19.36
	0.75	38.73 ± 18.59	38.54 ± 19.07	38.57 ± 18.75
Adaptive Thresholding (AT)	0.50	88.72 ± 12.90	88.04 ± 13.91	88.21 ± 13.02
	0.55	85.16 ± 14.36	84.51 ± 15.22	84.67 ± 14.46
	0.60	78.45 ± 15.65	77.87 ± 16.36	78.02 ± 15.75
	0.65	67.70 ± 17.08	67.25 ± 17.72	67.35 ± 17.19
	0.70	55.11 ± 18.30	54.80 ± 18.85	54.86 ± 18.42
	0.75	39.50 ± 18.16	39.33 ± 18.41	39.35 ± 18.19
Template Matching (TM)	0.50	63.90 ± 37.38	65.35 ± 37.11	64.40 ± 37.06
	0.55	60.79 ± 36.86	62.10 ± 36.44	61.24 ± 36.47
	0.60	55.89 ± 35.20	56.99 ± 34.83	56.27 ± 34.85
	0.65	49.61 ± 32.83	50.48 ± 32.52	49.90 ± 32.54
	0.70	40.66 ± 28.60	41.25 ± 28.34	40.85 ± 28.36
	0.75	30.51 ± 23.14	30.94 ± 22.95	30.64 ± 22.96
AT Mask Prediction (AT-pred)	0.50	86.29 ± 14.81	86.79 ± 15.08	86.18 ± 14.48
	0.55	82.80 ± 15.67	83.45 ± 16.15	82.80 ± 15.60
	0.60	76.25 ± 17.05	76.74 ± 17.32	76.21 ± 16.98
	0.65	65.19 ± 18.10	65.72 ± 18.56	65.21 ± 18.14
	0.70	51.89 ± 17.22	52.46 ± 17.95	51.98 ± 17.44
	0.75	35.12 ± 16.72	35.48 ± 17.34	35.17 ± 16.91

Figure 19 shows the same precision, recall and F1-scores as Table 2 plotted against their respective IoU thresholds. Additionally, this figure shows a histogram of the relative counts each F1-score is measured for IoU ≥ 0.5. For AT and AT-pred the F1-score drop below 0.80 when the IoU threshold becomes greater than 0.55. For all methods, we see that the precision, recall and F1-scores drop-off significantly when the IoU threshold becomes greater than 0.6. The histogram of AT shows that most F1-scores seem to accumulate above 0.80 and it shows the peak value at being in the 0.99-1.00 range when IoU ≥ 0.5. The same is true for AT-pred. For template, matching however, we stull see some accumulation above the 0.80 range but the peak value is instead in the 0.00-0.01 range.

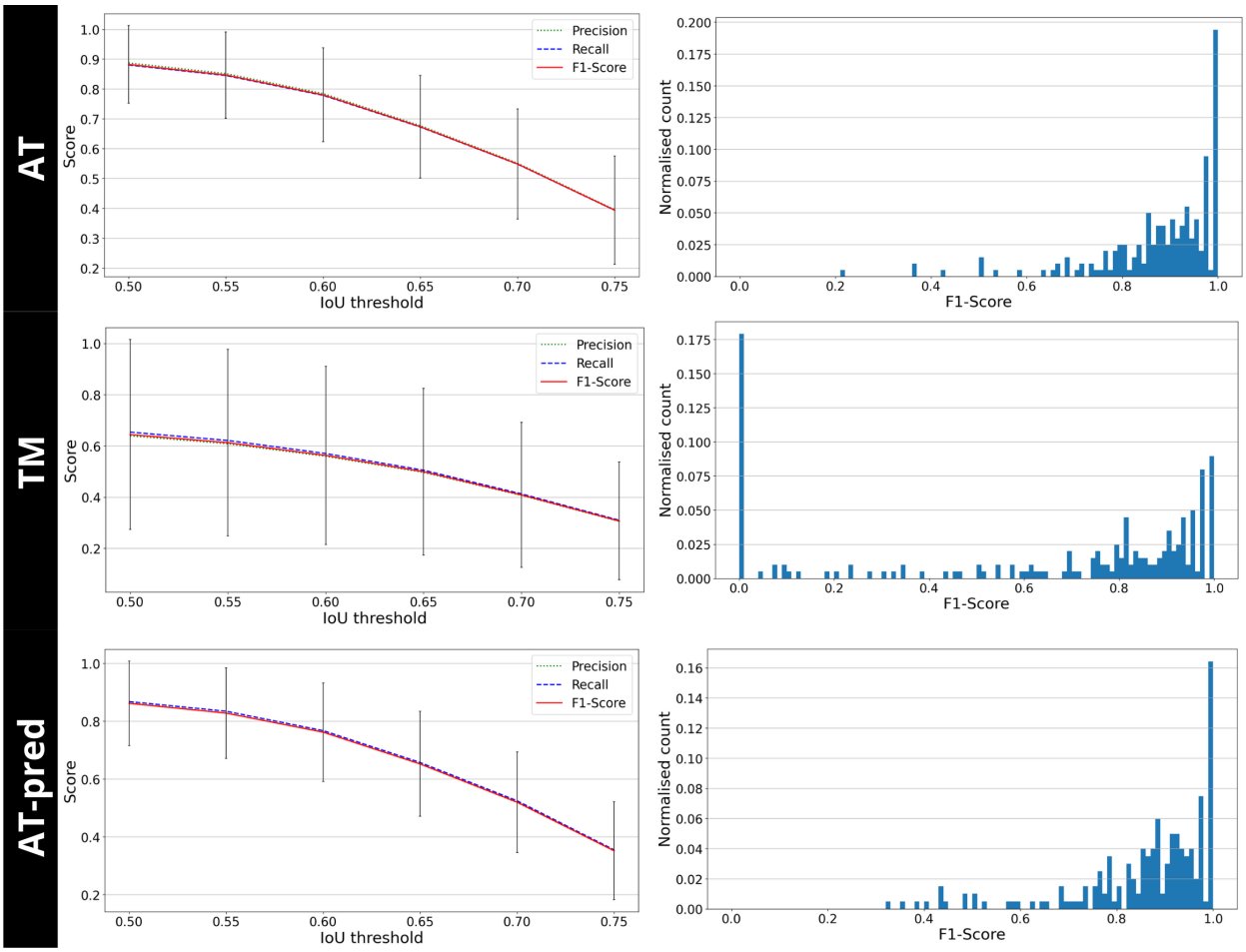


Figure 19: Column 1 shows the precision, recall and F1-score plotted against different IoU thresholds with the standard deviation plotted for the F1-scores for the adaptive threshold algorithm (AT), template matching algorithm (TM) and adaptive thresholding with mask predictions (AT-pred). Column 2 shows a histogram with the normalised counts of each F1-score for IoU ≥ 0.5 for each method.

The results of sensor localisation is shown for some example frames in Figure 20. In Figure 20a, we see an example of very good sensor localisation with an F1-score of 1.0 for both adaptive thresholding and template matching when IoU ≥ 0.5. Figure 20b contains some mistakes for both algorithms: AT misses one sensor and TM has some misaligned sensors. Sensors are also being detected outside the region of interest in both cases. Regardless performance is still relatively good with an F1-score of 0.82 for AT and 0.70 for TM when IoU ≥ 0.5. In Figure 20c, we see that, while the rest of the sensors are identified correctly, the first sensor is misaligned for AT resulting in an F1-score of 0.95 when IoU ≥ 0.5. This error is not present in TM but the F1-score is still 0.95 when IoU ≥ 0.5.

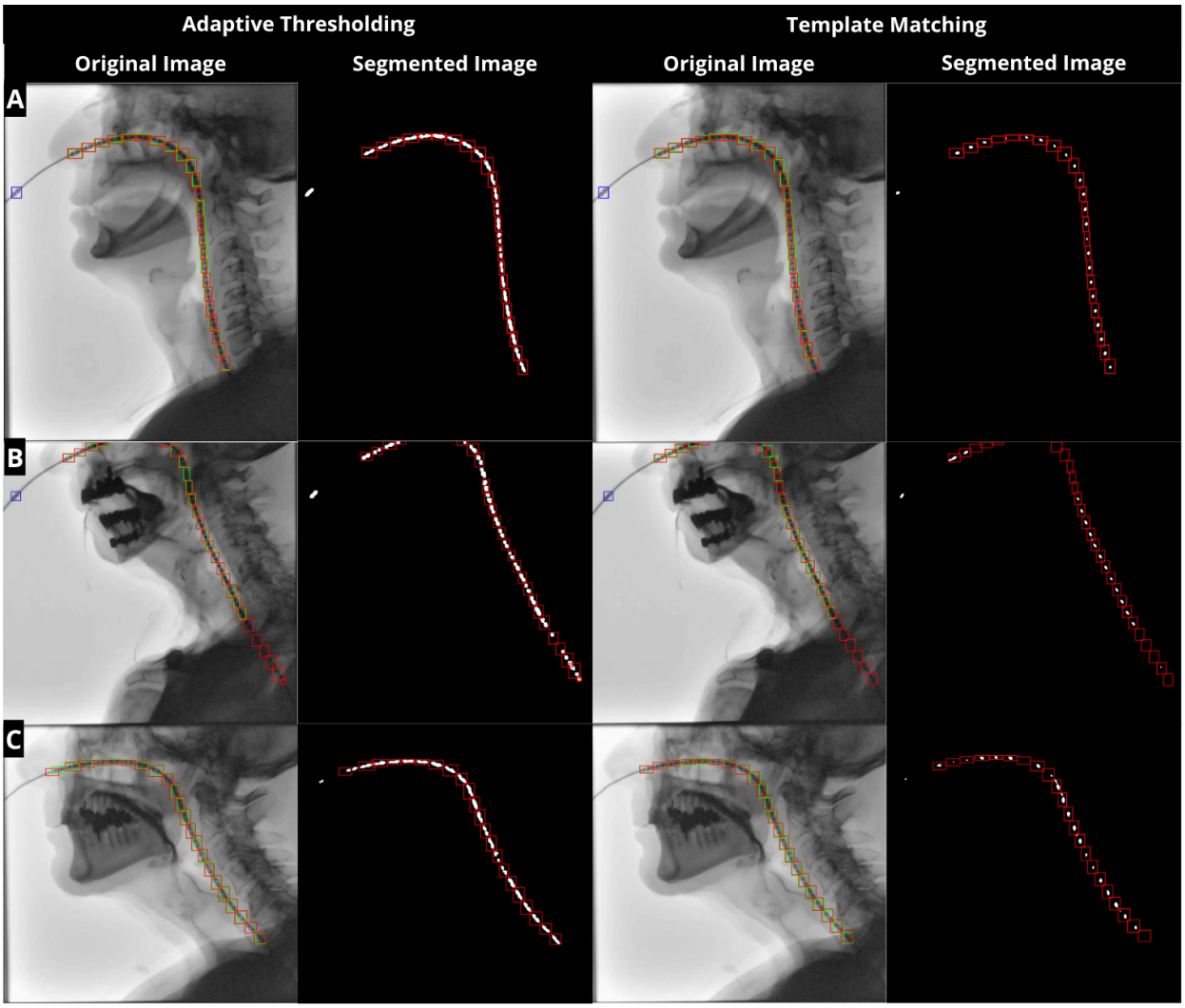


Figure 20: Some examples of segmented images. The first two columns show the results for adaptive thresholding and the latter two columns the results for template matching. The first column of both methods show the original VFSS frame with ground truth bounding boxes in green, the detected sensors in red and the reference sensor, which is part of the ground truth, in blue.

Some more examples of sensor localisation is shown in Figure 21, where the performance of the different algorithms is shown under different conditions. Figure 21a shows the sensor localisation in a frame with high levels of motion blur, Figure 21b shows the localisation for a frame where the bolus obstructs a large part of the HRIM catheter and Figure 21c shows the localisation for a frame where a foreign object obstructs a part of the catheter. In Figure 21a, the adaptive thresholding algorithm makes a minor mistake around a part of the catheter where there is some bolus residue but this mistake is not present when using the mask predictions or template matching. AT does detect some false positives, however. All algorithms seem to struggle at the around the bolus in Figure 21b but the other sensors in this frame are mostly identified correctly by AT and AT-pred. In Figure 21c, some errors are present but none near the implant.

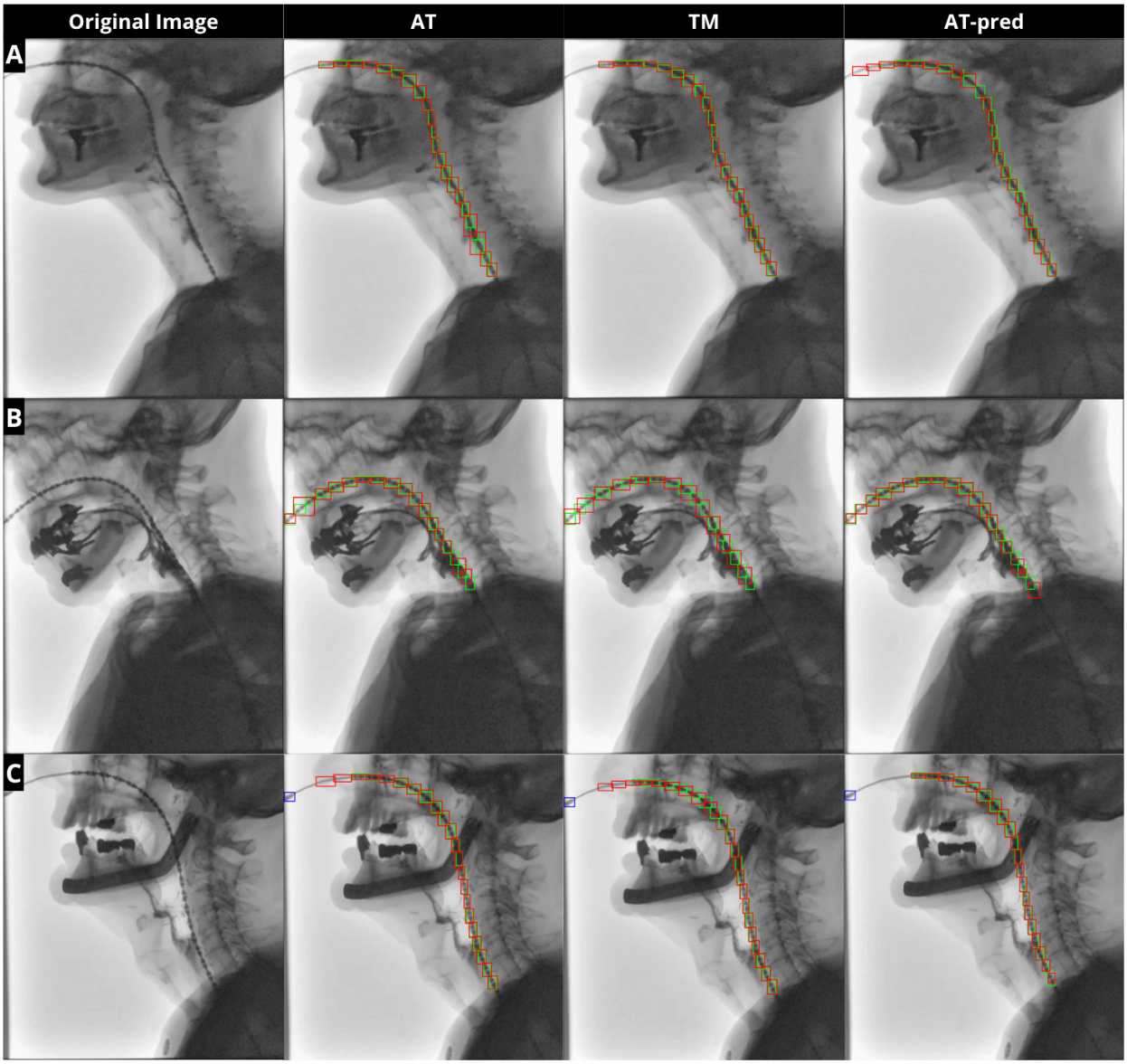


Figure 21: Some examples showing the sensor localisation of the algorithm using adaptive thresholding (AT), template matching (TM), and AT with mask predictions (AT-pred) under different different conditions. A) A frame with heavy motion blur artifacts. B) A frame where a the bolus obscures a large part of the catheter. C) A frame where a jaw implant obscures part of the catheter.

5 Discussion

The objective of this study was to develop an algorithm capable of reliably locating HRIM sensors from a VFSS video frame. The AT algorithm alone managed to achieve an F1-score of 88.21 ± 13.02 when IoU $\geq$ 0.5 (see Table 2) which indicates that the outlined algorithm can locate these sensors a majority of the time. Furthermore we tested this same algorithm using the predictions of the catheter mask from the paper by Rocha et al. [30] instead of the ground truth mask and it still performed roughly the same with an F1-score of 86.18 ± 14.48, meaning that the algorithm still performs well when this mask is imperfect. These results are comparable to those achieved by Geiger et al. [19], who used an algorithm based on template matching to localise sensors. We also tested an approach using template matching which uses templates extracted using the adaptive thresholding algorithm. This approach yielded much lower results, however. Achieving an F1-score of merely 65.35 ± 37.11 when IoU ≥ 0.5 (Table 2). A major contributor to this is the fact that no sensors are being detected at all in most frames (see Figure 19). This is most likely caused by a combination of an overly simple sensor detection method and a flawed scoring system (see Section 3.2.4).

The scoring system that is in place now gets rid of all or too many sensors for most frames making inter/extrapolation impossible or unreliable, due to the increased number of sensors that need to be inter/extrapolated. Furthermore, the method for obtaining the points used for this inter/extrapolation is flawed as well. The current algorithm uses a simple threshold over the normalised result matrix after template matching. This works alright in some cases but fails in others (see Figure 20) but we believe an approach similar to that used by Geiger et al. [19] would work better: They detect the highest intensity point and look for the next highest point within a certain distance. We suggest an approach that follows the same basic principle: Look for the highest intensity point and look for a local maximum around the same distance away as the estimated sensor length. This could be further enhanced by weighting these points by their distance to points found in the adaptive thresholding phase of the algorithm.

That being said, the adaptive thresholding step also has room for improvement. The backbone of the entire algorithm are the adaptive thresholding and the following morphological operations. The current algorithm applies morphological operations globally on all segments but an approach that applies opening/closing on segments individually until they are the right size, based on the estimated sensor length. This could prevent sensors splitting in half, merging with other sensors or being removed entirely. This would also ease the reliance on the splitting/merging step of the AT algorithm, which is one of the main points of error during this step. Namely problems occur when a short (half) segment is followed by a large (merged) segment. This would imply that the other half of the segment is merged with the other segment but the current algorithm fails to account for this, ignoring the smaller segment and mistaking the larger segment for a single sensor.

These types of mistakes can also cause problems further down the catheter. The merging of segments during the adaptive thresholding step (see Section 3.2.2) only considers the length of the sensors and not their relative position to other sensors. This means that it is possible for the algorithm to merge the end of one sensor with the start of another when these types of mistakes occur. Implementing this adaptive way of executing morphological operations would also improve the length estimation algorithm since there would be fewer shorter or longer segments (cf Figure 14) that throw off the mean/median. This could also prevent cases like shown in Figure 20c where the first sensor is offset. This happens when the first sensor is split in the thresholded because the algorithm assumes this is the latter half of a sensor due to the large space without sensors before this first segment.

Although these issues should be resolved before any serious clinical application, the AT algorithm proves promising as a way to locate sensors even in suboptimal conditions such as high levels of motion blur (cf. Figure 21). The algorithm does struggle when large obstructions, such as those caused by the bolus, are present but it is expected that a similar template matching methodology to that of Geiger et al. [19] could help resolve this issue, as demonstrated by the authors. The AT algorithm is still important, since it allows for the extraction of templates from the image itself, allowing for sensor localisation in a wider variety of conditions, such as different HRIM measurement devices.

This research has tested the algorithm on multiple random frames from different videos but in a real-world context, we would need to locate sensors throughout an entire video. This would make it possible

to compare each frame to past and future frames to further improve sensor localisation. This research has also exclusively focussed on traditional segmentation techniques. Although research on the segmentation of individual HRIM sensors in VFSS videos is not widespread, the paper by Xi et al. [31] about detecting catheters and electrodes in X-rays of the heart demonstrates that locating sensors using deep-learning is a viable option and should be explored further. An approach using a YOLO or U-net network, for example, could prove successful [32, 33].

Still, the algorithm described in this study provides a promising way of HRIM sensor localisation in VFSS frames. Although there are still some issues, the algorithm is accurately able to detect the size of these sensors with both the ground truth mask and mask predictions with an average error of 1.43 $\pm$ 1.13 mm and 1.43 $\pm$ 1.12 mm respectively. This algorithm also is able to detect the location of the sensors very well with an average error in the centre of the sensors locations of 1.08 $\pm$ 0.58 mm for the ground truth mask and 1.39 $\pm$ 0.73 mm for the mask predictions. This proves that the algorithm described in this study could prove very successful if the problems described here are solved.

6 Conclusion

In this study we aimed to develop an algorithm to locate the individual sensors of a HRIM manometer from a VFSS video frame. We developed an algorithm based on adaptive thresholding and expanded on this same algorithm by implementing template matching. The adaptive thresholding algorithm was able to locate the sensors in a majority of cases (F1-score = 87.08 $\pm$ 15.55 for IoU $\geq$ 0.5). Although template matching seems to be a promising way to improve on this algorithm further, our approach to extracting sensor locations from the response map was proved to be flawed resulting in lower performance (F1-score = 63.28 $\pm$ 36.67 for IoU $\geq$ 0.5). A new way locating sensors from the template matching response map should be found and implemented. Numerous improvements should also be made to the adaptive thresholding algorithm. Most notably, a more adaptive way of executing morphological opening/closing should be implemented during this step. We also suggest comparing each frame with past and future frames to more accurately determine sensor locations. Furthermore we suggest that research be done in implementing a deep-learning based approach to sensor localisation. Still, the algorithm described here was able to accurately determine the sensor length with an average error of 1.43 $\pm$ 1.13 mm and the sensor centres with an average error of 1.08 $\pm$ 0.58 mm. This suggests that the algorithm described by this study has the potential to be very successful if the suggestions described are implemented.

References

[1] Laura Q.M. Chow. "Head and Neck Cancer". In: *New England Journal of Medicine* 382.1 (Jan. 2020). Ed. by Dan L. Longo, pp. 60–72. ISSN: 1533-4406. DOI: 10.1056/nejmra1715715. URL: http://dx.doi.org/10.1056/NEJMra1715715.

[2] Freddie Bray et al. "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 74.3 (Apr. 2024), pp. 229–263. ISSN: 1542-4863. DOI: 10.3322/caac.21834. URL: http://dx.doi.org/10.3322/caac.21834.

[3] Deepak Lakshmipathy, Melissa Allibone, and Karthik Rajasekaran. "Dysphagia in Head and Neck Cancer". In: *Otolaryngologic Clinics of North America* 57.4 (Aug. 2024), pp. 635–647. ISSN: 0030-6665. DOI: 10.1016/j.otc.2024.02.013. URL: http://dx.doi.org/10.1016/j.otc.2024.02.013.

[4] Iris Krebbers et al. "Affective Symptoms and Oropharyngeal Dysphagia in Head-and-Neck Cancer Patients: A Systematic Review". In: *Dysphagia* 38.1 (July 2022), pp. 127–144. ISSN: 1432-0460. DOI: 10.1007/s00455-022-10484-8. URL: http://dx.doi.org/10.1007/s00455-022-10484-8.

[5] Maria Giulia Cristofaro et al. "The health risks of dysphagia for patients with head and neck cancer: a multicentre prospective observational study". In: *Journal of Translational Medicine* 19.1 (Nov. 2021). ISSN: 1479-5876. DOI: 10.1186/s12967-021-03144-2. URL: http://dx.doi.org/10.1186/s12967-021-03144-2.

[6] M G Rugiu. "Role of videofluoroscopy in evaluation of neurologic dysphagia". en. In: *Acta Otorhinolaryngol. Ital.* 27.6 (Dec. 2007), pp. 306–316.

[7] Bonnie Martin-Harris and Bronwyn Jones. "The Videofluorographic Swallowing Study". In: *Physical Medicine and Rehabilitation Clinics of North America* 19.4 (Nov. 2008), pp. 769–785. ISSN: 1047-9651. DOI: 10.1016/j.pmr.2008.06.004. URL: http://dx.doi.org/10.1016/j.pmr.2008.06.004.

[8] Janet W. Lee et al. "Subjective Assessment of Videofluoroscopic Swallow Studies". In: *Otolaryngology–Head and Neck Surgery* 156.5 (Feb. 2017), pp. 901–905. ISSN: 1097-6817. DOI: 10.1177/0194599817691276. URL: http://dx.doi.org/10.1177/0194599817691276.

[9] Jong Taek Lee, Eunhee Park, and Tae-Du Jung. "Automatic Detection of the Pharyngeal Phase in Raw Videos for the Videofluoroscopic Swallowing Study Using Efficient Data Collection and 3D Convolutional Networks". In: *Sensors* 19.18 (Sept. 2019), p. 3873. ISSN: 1424-8220. DOI: 10.3390/s19183873. URL: http://dx.doi.org/10.3390/s19183873.

[10] M. Panebianco et al. "Dysphagia in neurological diseases: a literature review". In: *Neurological Sciences* 41.11 (June 2020), pp. 3067–3073. ISSN: 1590-3478. DOI: 10.1007/s10072-020-04495-2. URL: http://dx.doi.org/10.1007/s10072-020-04495-2.

[11] K. Helliwell et al. "The use of videofluoroscopy (VFS) and fibreoptic endoscopic evaluation of swallowing (FEES) in the investigation of oropharyngeal dysphagia in stroke patients: A narrative review". In: *Radiography* 29.2 (Mar. 2023), pp. 284–290. ISSN: 1078-8174. DOI: 10.1016/j.radi.2022.12.007. URL: http://dx.doi.org/10.1016/j.radi.2022.12.007.

[12] Alicia K. Vose et al. "A Survey of Clinician Decision Making When Identifying Swallowing Impairments and Determining Treatment". In: *Journal of Speech, Language, and Hearing Research* 61.11 (Nov. 2018), pp. 2735–2756. ISSN: 1558-9102. DOI: 10.1044/2018_jslhr-s-17-0212. URL: http://dx.doi.org/10.1044/2018_JSLHR-S-17-0212.

[13] H. TOHARA et al. "Inter- and intra-rater reliability in fibroptic endoscopic evaluation of swallowing: RELIABILITY IN SWALLOWING EVALUATION". In: *Journal of Oral Rehabilitation* 37.12 (Nov. 2010), pp. 884–891. ISSN: 0305-182X. DOI: 10.1111/j.1365-2842.2010.02116.x. URL: http://dx.doi.org/10.1111/j.1365-2842.2010.02116.x.

[14] M. M. Szczesniak et al. "Inter-rater reliability and validity of automated impedance manometry analysis and fluoroscopy in dysphagic patients after head and neck cancer radiotherapy". In: *Neurogastroenterology & Motility* 27.8 (May 2015), pp. 1183–1189. ISSN: 1365-2982. DOI: 10.1111/nmo.12610. URL: http://dx.doi.org/10.1111/nmo.12610.

[15] Sanith S. Cheriyan et al. "Clinical application of pharyngeal high-resolution manometry in Ear, Nose and Throat (ENT) practice". In: *Australian Journal of Otolaryngology* 6 (Apr. 2023). ISSN: 2616-2792. DOI: 10.21037/ajo-22-37. URL: http://dx.doi.org/10.21037/ajo-22-37.

[16] N. Rommel et al. "Automated impedance manometry analysis as a method to assess esophageal function". In: *Neurogastroenterology & Motility* 26.5 (Jan. 2014), pp. 636–645. ISSN: 1365-2982. DOI: 10.1111/nmo.12308. URL: http://dx.doi.org/10.1111/nmo.12308.

[17] Taher I. Omari et al. "High-Resolution Pharyngeal Manometry and Impedance: Protocols and Metrics—Recommendations of a High-Resolution Pharyngeal Manometry International Working Group". In: *Dysphagia* 35.2 (June 2019), pp. 281–295. ISSN: 1432-0460. DOI: 10.1007/s00455-019-10023-y. URL: http://dx.doi.org/10.1007/s00455-019-10023-y.

[18] Ju Seok Ryu, Donghwi Park, and Jin Young Kang. "Application and Interpretation of High-resolution Manometry for Pharyngeal Dysphagia". In: *Journal of Neurogastroenterology and Motility* 21.2 (Apr. 2015), pp. 283–287. ISSN: 2093-0887. DOI: 10.5056/15009. URL: http://dx.doi.org/10.5056/15009.

[19] Alexander Geiger et al. "Towards multimodal visualization of esophageal motility: fusion of manometry, impedance, and videofluoroscopic image sequences". In: *International Journal of Computer Assisted Radiology and Surgery* 20.4 (Oct. 2024), pp. 713–721. ISSN: 1861-6429. DOI: 10.1007/s11548-024-03265-1. URL: http://dx.doi.org/10.1007/s11548-024-03265-1.

[20] Ying Yu et al. "Techniques and Challenges of Image Segmentation: A Review". In: *Electronics* 12.5 (Mar. 2023), p. 1199. ISSN: 2079-9292. DOI: 10.3390/electronics12051199. URL: http://dx.doi.org/10.3390/electronics12051199.

[21] Yan Xu et al. "Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches". In: *Bioengineering* 11.10 (Oct. 2024), p. 1034. ISSN: 2306-5354. DOI: 10.3390/bioengineering11101034. URL: http://dx.doi.org/10.3390/bioengineering11101034.

[22] N Senthilkumaran and S Vaithegi. "Image Segmentation By Using Thresholding Techniques For Medical Images". In: *Computer Science &; Engineering: An International Journal* 6.1 (Feb. 2016), pp. 1–13. ISSN: 2231-3583. DOI: 10.5121/cseij.2016.6101. URL: http://dx.doi.org/10.5121/cseij.2016.6101.

[23] Mahreen Kiran et al. "Comparative analysis of segmentation techniques based on chest X-ray images". In: *Multimedia Tools and Applications* 79.13–14 (Mar. 2019), pp. 8483–8518. ISSN: 1573-7721. DOI: 10.1007/s11042-019-7348-3. URL: http://dx.doi.org/10.1007/s11042-019-7348-3.

[24] OpenCV Development Team. *OpenCV Documentation: AdaptiveThresholdTypes*. Accessed: 2025-06-23. 2025. URL: https://docs.opencv.org/4.x/d7/d1b/group__imgproc__misc.html#ga72b913f352e4a1b1b397736707afcde3.

[25] Ardeshir Goshtasby. "Template Matching in Rotated Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-7.3 (May 1985), pp. 338–344. ISSN: 0162-8828. DOI: 10.1109/tpami.1985.4767663. URL: http://dx.doi.org/10.1109/tpami.1985.4767663.

[26] Jingning Yu. "Based on Gaussian filter to improve the effect of the images in Gaussian noise and pepper noise". In: *Journal of Physics: Conference Series* 2580.1 (Sept. 2023), p. 012062. ISSN: 1742-6596. DOI: 10.1088/1742-6596/2580/1/012062. URL: http://dx.doi.org/10.1088/1742-6596/2580/1/012062.

[27] Karel Zuiderveld. "Contrast Limited Adaptive Histogram Equalization". In: *Graphics Gems*. Elsevier, 1994, pp. 474–485. ISBN: 9780123361561. DOI: 10.1016/b978-0-12-336156-1.50061-6. URL: http://dx.doi.org/10.1016/B978-0-12-336156-1.50061-6.

[28] Mojdeh Mehdizadeh, Kioumars Tavakoli Tafti, and Parisa Soltani. "Evaluation of histogram equalization and contrast limited adaptive histogram equalization effect on image quality and fractal dimensions of digital periapical radiographs". In: *Oral Radiology* 39.2 (Sept. 2022), pp. 418–424. ISSN: 1613-9674. DOI: 10.1007/s11282-022-00654-7. URL: http://dx.doi.org/10.1007/s11282-022-00654-7.

[29] Mohamed Amine Larhmam, Mohammed Benjelloun, and Saïd Mahmoudi. "Vertebra identification using template matching modelmp and K-means clustering". In: *International Journal of Computer Assisted Radiology and Surgery* 9.2 (July 2013), pp. 177–187. ISSN: 1861-6429. DOI: 10.1007/s11548-013-0927-2. URL: http://dx.doi.org/10.1007/s11548-013-0927-2.

[30] M. M. Rocha et al. "Deep-learning segmentation of manometer in video-fluoroscopy swallow studies of head and neck cancer patients". In: *CRAS 2025 Conference on New Technologies for Computer/Robot Assisted Surgery*. Submitted. 2025.

[31] Xi Lin et al. "Catheter Detection and Segmentation in X-ray Images via Multi-task Learning". In: *arXiv preprint arXiv:2503.02717* (2025). Submitted March 4, 2025. URL: https://arxiv.org/abs/2503.02717.

[32] Sara Hashemi et al. "Use of Yolo Detection for 3D Pose Tracking of Cardiac Catheters Using Bi-Plane Fluoroscopy". In: *AI* 5.2 (June 2024), pp. 887–897. ISSN: 2673-2688. DOI: 10.3390/ai5020044. URL: http://dx.doi.org/10.3390/ai5020044.

[33] Ina Vernikouskaya et al. "Cryo-balloon catheter localization in X-Ray fluoroscopy using U-net". In: *International Journal of Computer Assisted Radiology and Surgery* 16.8 (Apr. 2021), pp. 1255–1262. ISSN: 1861-6429. DOI: 10.1007/s11548-021-02366-5. URL: http://dx.doi.org/10.1007/s11548-021-02366-5.