A KERNEL DENSITY ESTIMATION BASED QUALITY METRIC FOR QUALITY ASSESSMENT OF OBSTETRIC ULTRASOUND VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Simplified ultrasound scanning protocols (sweeps) have been developed to reduce the high skill required to perform a regular obstetric ultrasound examination. However, without automated quality assessment of the video, the utility of such protocols in clinical practice is limited. An automated quality assessment algorithm is proposed that applies an object detector to detect fetal anatomies within ultrasound videos. Kernel density estimation is applied to the bounding box annotations to estimate a probability density function of certain bounding box properties such as the spatial and temporal position during the sweeps. This allows to quantify how well the spatio-temporal position of anatomies in a sweep agree with previously seen data as a quality metric. The new quality metric is compared to other metrics of quality such as the confidence of the object detector model.

1 INTRODUCTION

Obstetric ultrasound scanning is a vital part of antenatal care; as it allows us to accurately date a pregnancy, identify pregnancy risk factors, check the health of the fetus and much more Crino et al. (2013); Salomon et al. (2011; 2019). Although ultrasound is considered a cheap and portable imaging modality, there is still a shortage of availability of obstetric ultrasound scans in limited resource settings Ngoya et al. (2016); Mollura & Lungren (2014); Mollura et al. (2013). However these are the regions where they are most needed; with 94% of pregnancy and childbirth related deaths occurring in developing nations and 80% of these occurring in areas of high birthrate and limited access to healthcare Organization et al. (2019). A major limitation for the availability of ultrasound is the lack of trained sonographers and healthcare workers Maru et al. (2010); Darmstadt et al. (2009) [World Health Organization and United Nations Children's Fund. WHO/UNICEF joint database on SDG 3.1.2 Skilled Attendance at Birth.]

The acquisition of obstetric ultrasound requires a high level of training and expertise, with the sonographers iteratively repositioning the probe whilst viewing real time at the surrounding anatomy to capture the most informative imaging plane. Additional difficulties include the movement of the fetal head and variable fetal position in the uterus.

Thus, simplified scanning protocols have been published that don't rely on the intuition and experience of the sonographers, but where the probe is to scan along predefined linear sweeps across the body outlined solely by external body landmarks of the pregnant woman Abu-Rustum & Ziade (2017); Abuhamad et al. (2015); DeStigter et al. (2011). These scanning protocols (sweeps) allow for acquisition without highly trained sonographers, which can then be analysed by clinicians remotely with the advent of tele-medicine Toscano et al. (2021); Marini et al. (2021), or by the medical image analysis algorithms.

In practice with these new acquisition protocols, we cannot assume the user will be able to differentiate between a usable scan or a low quality non-informative one. Ultrasound scans are cheaply retaken during the same appointment, however retaking a scan after the patient has left is a large burden to the patient. Whilst trained clinicians can tell when to retake a video if it's sub-optimal or not "fit-for-purpose", computers are yet to be able to mimic this human capability. Thus, developing automated algorithms for ultrasound video quality assessment is an important research challenge to aid in the adoption of ultrasounds where trained sonographers are not always available.

This paper seeks to provide a quality assessment method for sweep data. The method utilises bounding box annotations of ultrasound sweep data, to train an anatomy detector model, and via kernel density estimation, quantifies how well the spatial and temporal position of bounding boxes fit in with the expected "typical" data.

2 Related Work

Image quality assessment in signal processing is a well researched topic with many different metrics proposed like PSNR and SSIM. These however are image-based (not video), fully referenced methods (requires an undistorted image) and mostly focus on compression losses Thung & Raveendran (2009); Hore & Ziou (2010).

No-referenced video quality assessment (NRVQA) literature presents models designed to quantify a specific type of image distortion such as blur Marziliano et al. (2002), ringing Feng & Allebach (2006), blockiness Wang et al. (2000) banding Wang et al. (2016); Tu et al. (2020) or noise Amer & Dubois (2005). Current NRVQA metrics such as Li et al. (2019); Mittal et al. (2012) rely on using natural scene statistics and models of the human visual system. Thus they are specific for natural images, and likely will not generalise well to ultrasound videos, as these contain many types of distortions that are beyond the normal range for natural images.

Quality assessment of ultrasound video must therefore be defined differently; it should be based off clinical usefulness. This field has less research. Wu et al. (2017) and Lin et al. (2019) propose quality assessment off anatomy specific criteria that can be separately assessed by individual networks; Wu et al. (2017) develops deep learning models to check things like: if the fetal stomach bubble appears full and salient with a clear boundary, and Lin et al. (2019) "the lateral sulcus must be clearly visible" and "the skull is in the middle of the ultrasound plane and larger than 2/3 of overall fan shape area" and other criteria.

The most related work is by Komatsu et al. (2021) who developed an object detector network to detect cardiac structures in fetal ultrasound scans and used the detection results to generate an "abnormality score" for the heart. The work uses a specific imaging plane of the heart, the three-vessel trachea view and four-chamber view, where it is expected to see a full set of specific anatomical substructures in every frame of these clips. Thus they simply generated an abnormality score between 0-1 which depends linearly to the number of anatomies and number of frames detected.

3 Method

This work uses bounding box annotations provided by an expert sonographer for a simplified ultrasound scan. The method firstly finds the distribution of the spatial and temporal position of the anatomies for this type of scan from many videos. For a new video, it then evaluates how well the spatio-temporal position of the anatomies of a new video fits in the distribution.

3.1 Data

The data was gathered as part of the Computer Assisted Low-cost Point of Care Ultrasound (CALOPUS) Project (UK Research Ethics Committee 18/WS/0051). A simplified ultrasound scanning protocol proposed by Abuhamad et al. (2015) was refined by Self et al. (2022) to contain 5 steps of which are shown in Fig. 1.

The scans were taken by an experienced sonographer, who also annotated the scans frame by frame with a bounding box around each of 11 possible anatomical structures (listed in Fig. 2). Of the five sweeps outlined in Fig. 1, this work only uses T-shaped sweep (step 1). For this work, only scans of gestational age between 18-23 weeks, and cephalic presentation are used.

This was to ensure the scans looked as homogeneous as possible. With different fetal presentations, we expect different locations of anatomies, and anatomies to show up at different times along the scan. The anatomies will also be imaged along different views, which affects the way the anatomy will appear thus the size and shape of the bounding boxes too. The gestational age cutoffs were chosen for similar reasons, but also because the majority of scans lied within these bounds. This left us with 45 ultrasound videos.

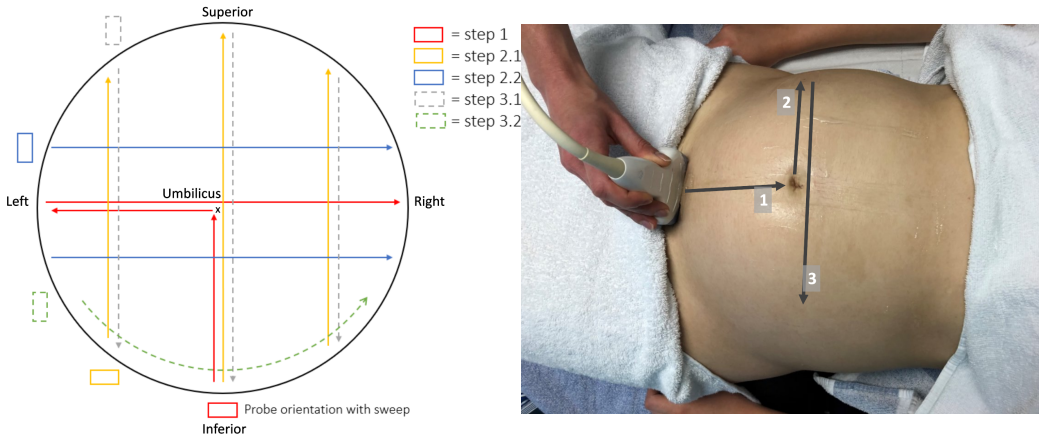

Figure 1: Left: all the different CALOPUS sweeps in Self et al. (2022) outlined in Section 3.1; right: the T-shaped sweep (step 1 on left) that we used in this work outlined on a participant.

Whilst we use only sweep 1 (see Fig. 1), this work could easily extend to any other sweeps. We chose the T-shaped sweep because it often showed the most anatomies, thus bounding boxes. The different bounding box annotations consisted of 11 anatomies, of which were distributed as shown in Fig. 2.

The CALOPUS project is a collaboration with the Translational Health Science and Technology Institute (THSTI) based in Delhi, India Self et al. (2022). THSTI have also scanned 72 patients using the same protocol, with cephalic presentation and of 18-23 gestational age, these cannot be viewed/accessed. However, the bounding box annotations of these scans which include: the bounding box co-ordinates, frame number and anatomy can be utilised. This gives an extra 72 videos' bounding box information.

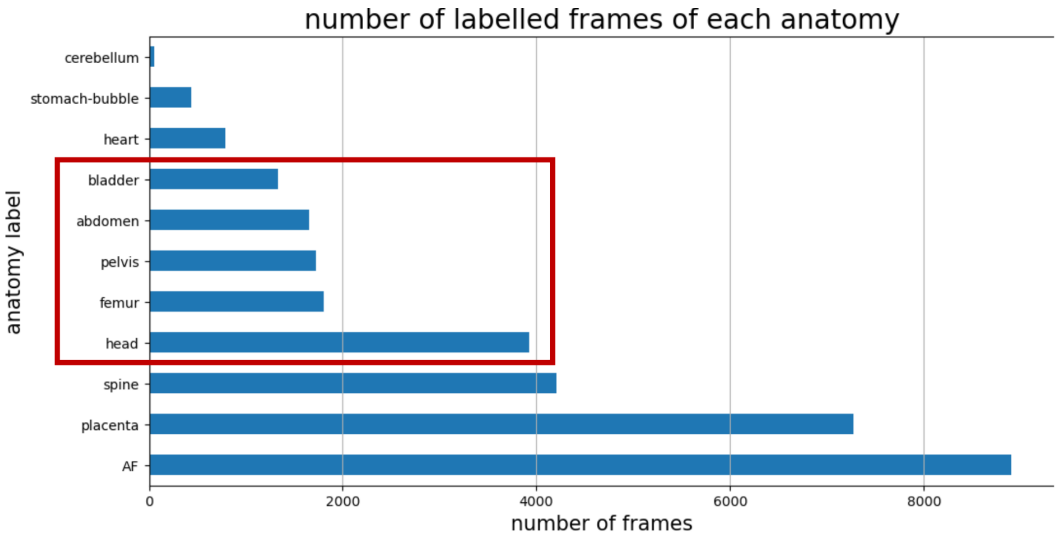

Figure 2: Number of frames per anatomy in the UK dataset after applying exclusion criteria (see Section 3.1) **Maybe make this so font is bigger, so can make fig smaller/ portrait?**.

3.2 Overview

A brief overview of our pipeline is: (1) using the bounding box annotated videos, we train an object detector that can produce bounding boxes around fetal anatomies. (2) Estimate the probability distribution of spatial position and timing of the bounding boxes for a cephalic T-shaped sweep using

kernel density estimation. (3) Perform inference on new data with our trained anatomy detector model to get bounding boxes for each anatomy for this video. (4) Compare the spatial and temporal position of the bounding boxes of the new video against our distribution of annotated videos. If the new bounding box properties fit in well with our estimated distribution, we propose it is of high quality, whilst if it doesn't, we propose it is abnormal or low quality. Numerical values for 'how well it fits the distribution' can be produced via calculating the probability of the bounding box having a property of a certain value or a less likely value via integration of the probability density function (PDF). One probability score for each bounding box (so frame-level), and averaging the probability over the entire video clip is our quality metric.

3.3 Anatomy Detector Model

Our anatomy detector model was based off the RetinaNet architecture with focal loss Lin et al. (2017). This is a one-shot detector, that contains feature pyramid networks to propagate multiple scale features down the network. The network has subnets for object classification and bounding box location regression. The backbone used was simply a ResNet-50 architecture He et al. (2016) that was pretrained off of ImageNet1k image classification dataset. The overall network was pretrained itself with the COCO dataset to achieve a mean average precision of the bounding box of 36.9 in the COCO dataset. This is a one-shot detector, and can do real-time inference (30 frames per second).

The data imbalance as seen in Fig. 2 caused large drops in performance; anatomies such as the cerebellum and stomach bubble were never being detected, whilst the placenta and AF were over predicted.

Thus we used a subset of only the 6 anatomies: the spine, head, femur, pelvis, abdomen and bladder. There is still a considerable data imbalance, but not as problematic as with all 12. These anatomies were chosen as they were consistently showing in the majority of the scans, and were not within each other - i.e. the stomach bubble will be within the abdomen, cerebellum is within the head.

3.4 Defining Bounding Box Properties

We chose to use the timing, spatial position of the bounding boxes as discriminators to determine if a scan was 'typical' or not. There were other properties that could have been used: the size, the aspect ratio, properties of the image content inside the bounding box like texture, gradients and more. Size, and aspect ratio were found to be uninformative, because the bounding box annotations were not tightly aligned to the exact edge of the anatomy structure, but often loosely drawn to the approximate size and location of the anatomy - thus seemed uninformative of the actual shape and size of the anatomy visible.

3.5 Probability Density Function of Bounding Box Properties

We use kernel density estimation (KDE) with a Gaussian kernel to estimate the probability distribution function (PDF) of the timing and position of the bounding boxes. A plot of the PDF and histogram for the timing of the femur is shown in Fig 3.

However we cannot assume these properties of bounding boxes are independent, rather, it is likely they are highly dependant. For example, due to the shape of the sweep, the timing and the position of the bounding boxes interact; as we pan side to side (the top of the "T" of the sweep), the anatomies first go right, then left in that order, thus the x co-ordinate and the time of the bounding boxes are highly correlated. We account for these dependencies by, instead of having a probability distribution function for each property, having a joint multi-dimensional probability distribution for all properties. A two-dimensional probability distribution of the x and y co-ordinate of the head bounding boxes is shown in Fig. 3

We extend this to a three dimensional joint probability distribution, with the axes as: x, y-co-ordinate and time. This is a three dimensional function, so cannot be visualised easily, but plotting this on top of a frame at different time points gives Fig. 4.

The bandwidth of the kernel used in the kernel density estimation strongly influences the overall PDF - much more so than the shape of the kernel (Turlach et al., 1993). In this work we use a Gaussian kernel, and fine-tuned the bandwidth using contextual knowledge. Both Silverman's rule

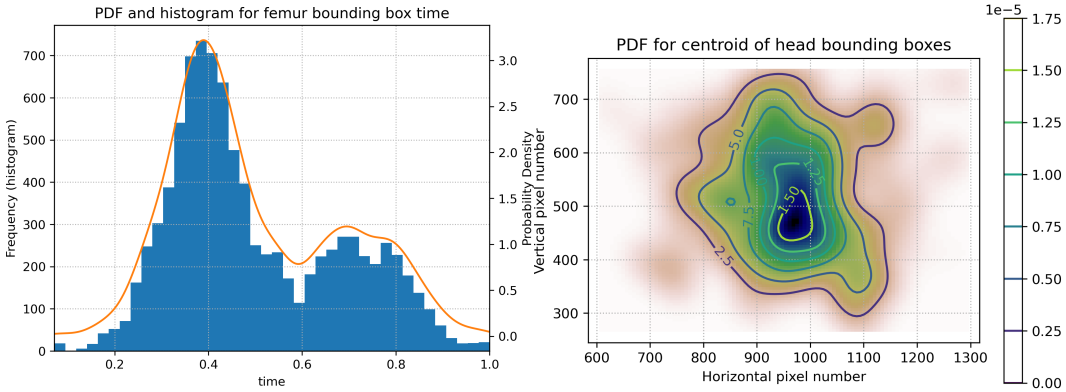

Figure 3: Left: histogram and PDF of the timing for the femur; right: the joint PDF for x, y coordinates of the head bounding boxes.

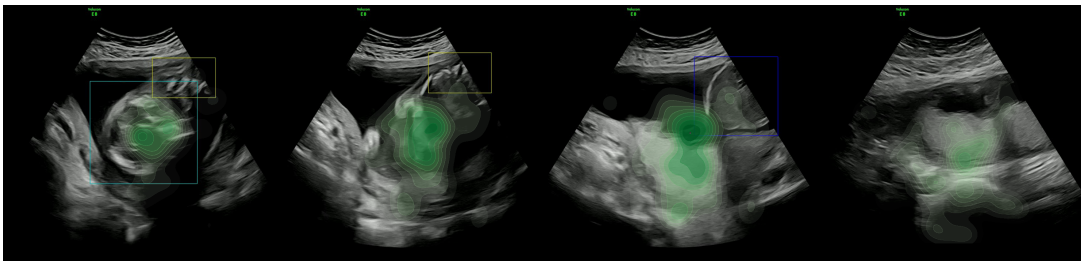

Figure 4: The change across time for the contour of the PDF for the spatial position of the head in the context of the actual ultrasound frame. From left to right, time point in the scan increases. The more intensity of the green reflects the higher PDF values.

and Scott's rule gave very similar PDFs, however the PDF was very peaky and multi-modal as can be seen in Fig. 5. The rightmost contour plot shows steep decreases in the PDF value within a 20 pixel radius of the mode, and many different peaks. It would be wrong to assume a spine that appears 20 pixels away from this peak indicates a much lower quality scan. We believe there are two general regions within the fan shaped area of the ultrasound where we expect to find a spine during the scan, so we adjust the bandwidth to reflect this. With too large bandwidth, we lose spatial resolution. Similar intuition was used for the temporal domain too.

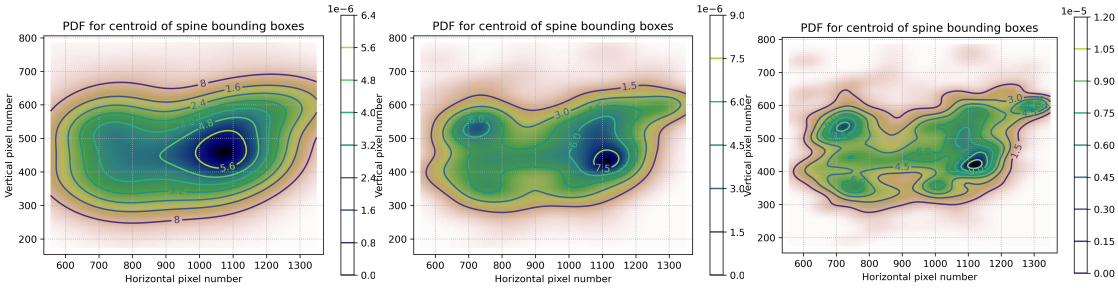

Figure 5: Contour plots of the PDF for the spatial position of the spine bounding boxes. The PDF becomes less peaky as the kernel bandwidth decreases from left to right. The right PDF was produced when using Scott's Rule for bandwidth selection.

3.6 Integration of the Probability Density Function

To get a numerical value of how well the position and timing of the bounding box fits in with the estimated distribution, we can find the probability of the bounding box having this position and

timing or a less likely timing/position (equivalent to a p-value). Thus we integrate the PDF for all areas where the PDF evaluates to a lower probability density than that of a given position and timing. I.e.

$$p = \iiint_V f(x, y, t)\, dx\, dy\, dt$$

where f is the PDF, x, y, t are the x and y co-ordinates, and time of the bounding box respectively, and the limits of integration V is the volume/region in x, y, t domain that encloses:

$$V = V(x, y, t) \quad \text{where:} \quad f(x, y, t) < f(x_{bbox}, y_{bbox}, t_{bbox})$$

Rather than direct integration of the probability density function to find the probability, we used Monte Carlo sampling methods to estimate the probability that the integral evaluates to. The exact algorithm is written in pseudo-code in the Appendix.

This method relies on using many random samples of the PDF to get an estimate of the PDF, thus becomes increasingly more accurate and precise with more samples. Although sampling many times takes time, it takes much less time than using the built-in integration libraries available in sci-py as we don't require such exact solutions. With this method, we can change the precision with the number of samples we use. As we require one integration with each bounding box, and there often being more than one bounding box per frame of a video of thirty seconds, we compromise between run time and precision. A plot of precision and number of samples for the same integral is shown in Fig. 6.

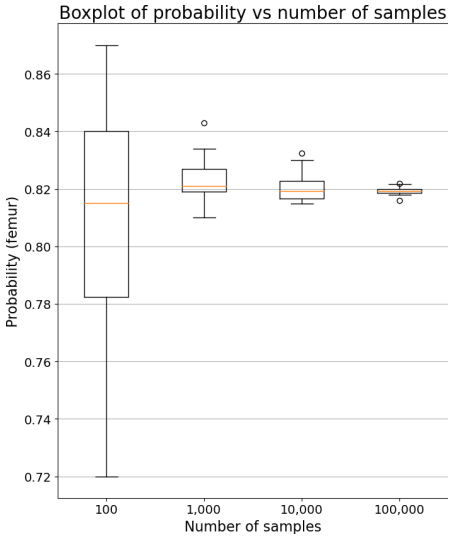

Figure 6: Box plot of a random integral (an arbitrary femur position and time) performed 30 times, but with increasing sample number for the Monte-Carlo method.

Whilst the time taken for each integration is directly proportional to the sample size, the error is inversely proportional to the square root of the sample size, thus we compromised with 1,000 samples for each integral.

To get from probabilities to a quality score, we took the mean probability for the video. Therefore our quality score is a measure of how similar are the bounding boxes of the new scan to all our current annotated scans that we have used to generate the PDF. We thus assume that a "high quality scan" is one which is similar to all the scans we have annotated. Almost all our current annotated scans were seen as highly typical and easily annotated, so this assumption holds.

Our quality metric uses only bounding box properties without explicitly requiring manual annotations of quality score by an expert. A large motivation for this method of quality assessment, was to leverage our already annotated dataset rather than requiring new annotations, something we imagine other groups that apply computer vision to medical imaging can use.

Additionally, basing a quality score off the bounding boxes makes sense as these bounding boxes contain the clinically important features in the video, and the rest outside the bounding boxes aren't focused on by the sonographer.

4 Experiments and Results

4.1 Anatomy Detector Model

To train the anatomy detector model, the data was split at the patient level, with 31:7:7 patients respectively for training, validation and testing. Simple data augmentation was used: random horizontal flips, brightness, crops, and slight rotations. The minority classes were not over-sampled or over-augmented, but proportionally sampled. The network was trained for 80 epochs, with an initial learning rate: 0.001, which dropped to 0.0001 at 40 epochs. The batch size was 16, and momentum: 0.8. Early stopping was used, so the best validation model saved. The trained model achieved the results as shown in the appendix.

We can assess how well the detector performs for our purpose by comparing it with a "perfect" detector (the ground truth annotations). A scatter plot of the probability for each bounding box against time can be used for this comparison. If the plots look similar, and mean probability values are similar then the detector is good enough. The comparison is shown in Fig. 7

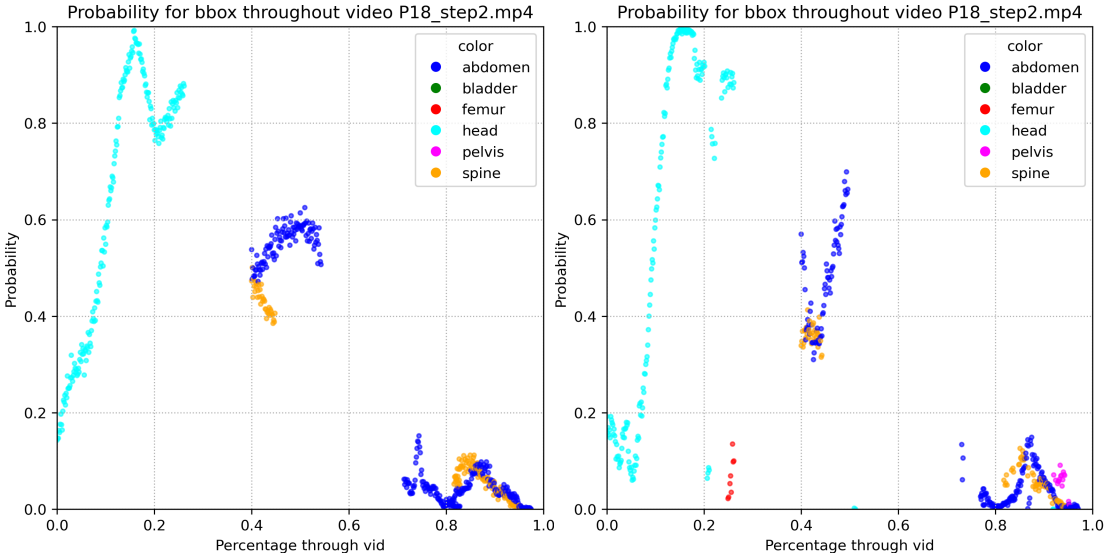

Figure 7: The vertical axes show how well the spatial and temporal position of the anatomy fits with our seen data. The higher up on the vertical axes, the better the fit with our distribution from the PDF. Each point in the plot is a bounding box of an anatomy.

Table 1: Table of the mean and standard deviation of the probability score for our test set videos for the model vs ground truth.

Video	Mean (GT)	Mean (Model)	Std. Dev (GT)	Std. Dev (Model)
P144_step1	0.453	0.450	0.279	0.236
P163_step1	0.406	0.274	0.310	0.196
P166_step1	0.466	0.515	0.284	0.279
P115_step1	0.308	0.372	0.298	0.297
P18_step1	0.322	0.301	0.313	0.331

4.2 PROBABILITY MODEL AS A METRIC FOR QUALITY

In this work, a high quality scan is one where the anatomies appear spatially and temporally at a similar position along the scan as the previous scans. The videos have not been annotated with a quality score since they are all typically of high quality as part of the data gathering process. Therefore we use other types of sweeps (see Fig. 1) and other fetal presentations as our ”bad quality” scan. The anatomies should appear at different locations and timings for different fetal presentations/scans, thus will serve as our ”bad quality” scan and our method should differentiate between these.

Thus we run our method on 7 videos of step 1, but breech presentation, and 3 videos each of step 2.1, 2.2, 3.1 (Fig. 1) each, (all in cephalic presentation).

We expect the step 1 cephalic sweeps to have the highest mean probability score, and the breech and other steps to have lower ones. The mean probability should be noticeably higher in the step 1 cephalic than any others (similar to one class classification). A box plot of the probability scores are shown on the left of Fig. 8.

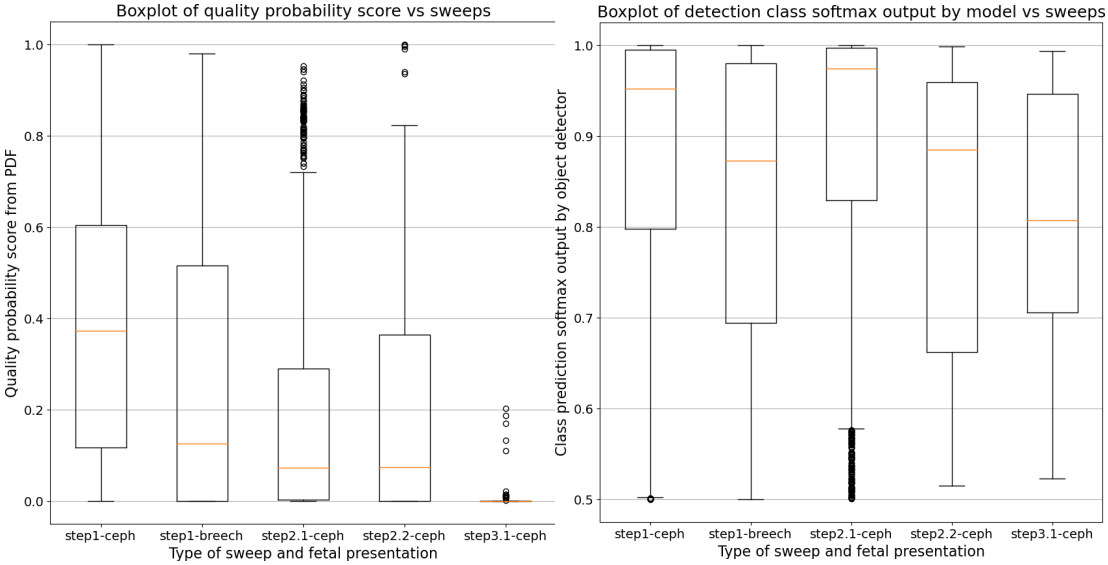

Figure 8: The left is a box plot using the ground truth bounding box annotations for the videos. And the right shows the quality score using the model’s detected bounding box . Step-1-ceph means step 1 in Fig. 1 with a cephalic presentation fetus.

To have a comparison we use the softmax output by the detection model for the class of the object. From intuition, if the detector sees a head that is very similar to all the previous heads it has been trained on, then the model will confidently predict that structure as a head, and so the softmax output for the head class for this bounding box will be almost 1. However some head bounding boxes will have much lower values, which will arise from the structure looking less similar to the training samples. So with different types of sweeps and different fetal presentations, the ultrasound view slices the anatomy differently, so the anatomies look visibly different, thus we expect much lower class confidence for these other sweeps and presentations.

We can see from Fig. 8 that there is no obvious difference between the mean values of the step 1 cephalic sweeps vs any of the others using this class confidence method.

5 CONCLUSION

We present a method of using kernel density estimation to assess whether the spatial and temporal position of anatomies in a specific scan follow the typical distribution. We show that this is an effective method for discriminating between different sweep steps and fetal presentations. We compare this to using the detector class softmax value as a way of discriminating between the different sweep steps and presentations, however we find our kernel density estimation method works much better.

REFERENCES

Reem S Abu-Rustum and M Fouad Ziade. The 3-sweep approach: A standardized technique for fetal anatomic assessment in the limited resource setting. *Journal of Fetal Medicine*, 4(1):25–30, 2017.

Alfred Abuhamad, Yili Zhao, Sharon Abuhamad, Elena Sinkovskaya, Rashmi Rao, Camille Kanaan, and Lawrence Platt. Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination. *American Journal of Perinatology*, 33:90–98, 8 2015. ISSN 10988785. doi: 10.1055/s-0035-1558828.

Aishy Amer and Eric Dubois. Fast and reliable structure-oriented video noise estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):113–118, 2005.

Jude Crino, Harris J Finberg, Faith Frieden, Jeffrey Kuller, Anthony Odibo, Alfred Robichaux, Marcela Bohm-Velez, Dolores H Pretorius, Sheila Sheth, Teresita L Angtuaco, et al. Aium practice guideline for the performance of obstetric ultrasound examinations. *Journal of Ultrasound in Medicine*, 32(6):1083–1101, 2013.

Gary L Darmstadt, Anne CC Lee, Simon Cousens, Lynn Sibley, Zulfiqar A Bhutta, France Donnay, Dave Osrin, Abhay Bang, Vishwajeet Kumar, Steven N Wall, et al. 60 million non-facility births: who can deliver in community settings to reduce intrapartum-related deaths? *International Journal of Gynecology & Obstetrics*, 107:S89–S112, 2009.

Kristen K DeStigter, G Eli Morey, Brian S Garra, Matthew R Rielly, Martin E Anderson, Michael G Kawooya, Alphonsus Matovu, and Frank R Miele. Low-cost teleradiology for rural ultrasound. In *2011 IEEE Global Humanitarian Technology Conference*, pp. 290–295. IEEE, 2011.

Xiaojun Feng and Jan P Allebach. Measurement of ringing artifacts in jpeg images. In *Digital Publishing*, volume 6076, pp. 74–83. SPIE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Masaaki Komatsu, Akira Sakai, Reina Komatsu, Ryu Matsuoka, Suguru Yasutomi, Kanto Shozu, Ai Dozen, Hidenori Machino, Hirokazu Hidaka, Tatsuya Arakaki, et al. Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning. *Applied Sciences*, 11(1): 371, 2021.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2351–2359, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zehui Lin, Shengli Li, Dong Ni, Yimei Liao, Huaxuan Wen, Jie Du, Siping Chen, Tianfu Wang, and Baiying Lei. Multi-task learning for quality assessment of fetal head ultrasound images. *Medical image analysis*, 58:101548, 2019.

Thomas J Marini, Daniel C Oppenheimer, Timothy M Baran, Deborah J Rubens, Marika Toscano, Kathryn Drennan, Brian Garra, Frank R Miele, Gail Garra, Sylvia Jacobo Noone, et al. New ultrasound telediagnostic system for low-resource areas: Pilot results from peru. *Journal of Ultrasound in Medicine*, 40(3):583–595, 2021.

Duncan Smith-Rohrberg Maru, Ryan Schwarz, Jason Andrews, Sanjay Basu, Aditya Sharma, and Christopher Moore. Turning a blind eye: the mobilization of radiology services in resource-poor regions. *Globalization and health*, 6(1):1–8, 2010.

Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. A no-reference perceptual blur metric. In *Proceedings. International conference on image processing*, volume 3, pp. III–III. IEEE, 2002.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Daniel Mollura and Matthew P Lungren. *Radiology in global health*, volume 1. Springer, 2014.

Daniel J Mollura, Jonathan Mazal, Kathryn L Everton, and RAD-AID Conference Writing Group. White paper report of the 2012 rad-aid conference on international radiology for developing countries: planning the implementation of global radiology. *Journal of the American College of Radiology*, 10(8):618–624, 2013.

Patrick Sitati Ngoya, Wilbroad Edward Muhogora, and Richard Denys Pitcher. Defining the diagnostic divide: an analysis of registered radiological equipment resources in a low-income african country. *The Pan African Medical Journal*, 25, 2016.

World Health Organization et al. Trends in maternal mortality 2000 to 2017: estimates by who, unicef, unfpa, world bank group and the united nations population division. 2019.

L. J. Salomon, Z. Alfirevic, F. Da Silva Costa, R. L. Deter, F. Figueras, T. Ghi, P. Glanc, A. Khalil, W. Lee, R. Napolitano, A. Papageorghiou, A. Sotiradis, J. Stirnemann, A. Toi, and G. Yeo. Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics and Gynecology*, 53:715–723, 6 2019. ISSN 14690705. doi: 10.1002/uog.20272.

Laurent Julien Salomon, Z Alfirevic, V Berghella, C Bilardo, E Hernandez-Andrade, SL Johnsen, K Kalache, K-Y Leung, G Malinger, H Munoz, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics & Gynecology*, 37(1): 116–126, 2011.

A Self, Q Chen, JA Noble, AT Papageorghiou, Bapu Koundinya Desiraju, Sumeet Dhariwal, Alexander Gleed, Divyanshu Mishra, Ramachandran Thiruvengadam, Varun Chandramohan, Rachel Craik, Elizabeth Wilden, and Ashok Khurana. Developing clinical artificial intelligence for obstetric ultrasound to improve access in underserved regions: the computer-assisted low-cost point-of-care ultrasound (calopus) study protocol. *Journal of Medical Internet Research*, 2022. doi: 10.2196/37374. URL `https://www.researchprotocols.org/2022/0/e0/`.

Kim-Han Thung and Paramesran Raveendran. A survey of image quality measures. In *2009 international conference for technical postgraduates (TECHPOS)*, pp. 1–4. IEEE, 2009.

Marika Toscano, Thomas J. Marini, Kathryn Drennan, Timothy M. Baran, Jonah Kan, Brian Garra, Ann M. Dozier, Rafael L. Ortega, Rosemary A. Quinn, Yu T. Zhao, Miguel S. Egoavil, Lorena Tamayo, Claudia Carlotto, and Benjamin Castaneda. Testing telediagnostic obstetric ultrasound in peru: a new horizon in expanding access to prenatal ultrasound. *BMC Pregnancy and Childbirth*, 21, 12 2021. ISSN 14712393. doi: 10.1186/s12884-021-03720-w.

Zhengzhong Tu, Jessie Lin, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Adaptive debanding filter. *IEEE Signal Processing Letters*, 27:1715–1719, 2020.

Berwin A Turlach et al. Bandwidth selection in kernel density estimation: a rewiew. Technical report, Humboldt Universitaet Berlin, 1993.

Yilin Wang, Sang-Uok Kum, Chao Chen, and Anil Kokaram. A perceptual visibility metric for banding artifacts. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2067–2071. IEEE, 2016.

Zhou Wang, Alan C Bovik, and Brian L Evan. Blind measurement of blocking artifacts in images. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 3, pp. 981–984. Ieee, 2000.

Lingyun Wu, Jie-Zhi Cheng, Shengli Li, Baiying Lei, Tianfu Wang, and Dong Ni. Fuiqa: fetal ultrasound image quality assessment with deep convolutional networks. *IEEE transactions on cybernetics*, 47(5):1336–1349, 2017.

A APPENDIX

Pseudo-code of the Monte-Carlo sampling used to estimate the probability

```
# We want to find the area under the pdf
for all x,y,s,t where the pdf evaluates to
a value lower than our bounding box p_density.

function pdf(x,y,t) #input shape:  (3,)
return prob_density   #output shape: (1,)

# is a function returns the probability
density for the pdf evaluated at (x,y,t).

# Evaluate the pdf for new bounding box
x,y,t:
p_density_of_bbox = pdf(x_bbox,y_bbox,t_bbox)

# Randomly sample from your pdf 10000 times
to get a list of x,y,t data:
samples_xyt = sample(pdf, size=10000)   # shape = (3,10000)

# The p of the sample evaluating to below
our p_density_bbox = value of the integral.

# So we count the number of samples where
p_density_sample < p_density_bbox and divide
by the number of all the samples, to get an
estimate of this probability:
lowsamples = pdf(samples_xyst) < p_density_of_bbox
# lowsamples is a 10000 logical vector
integral = sum(lowsamples) / 10000
```

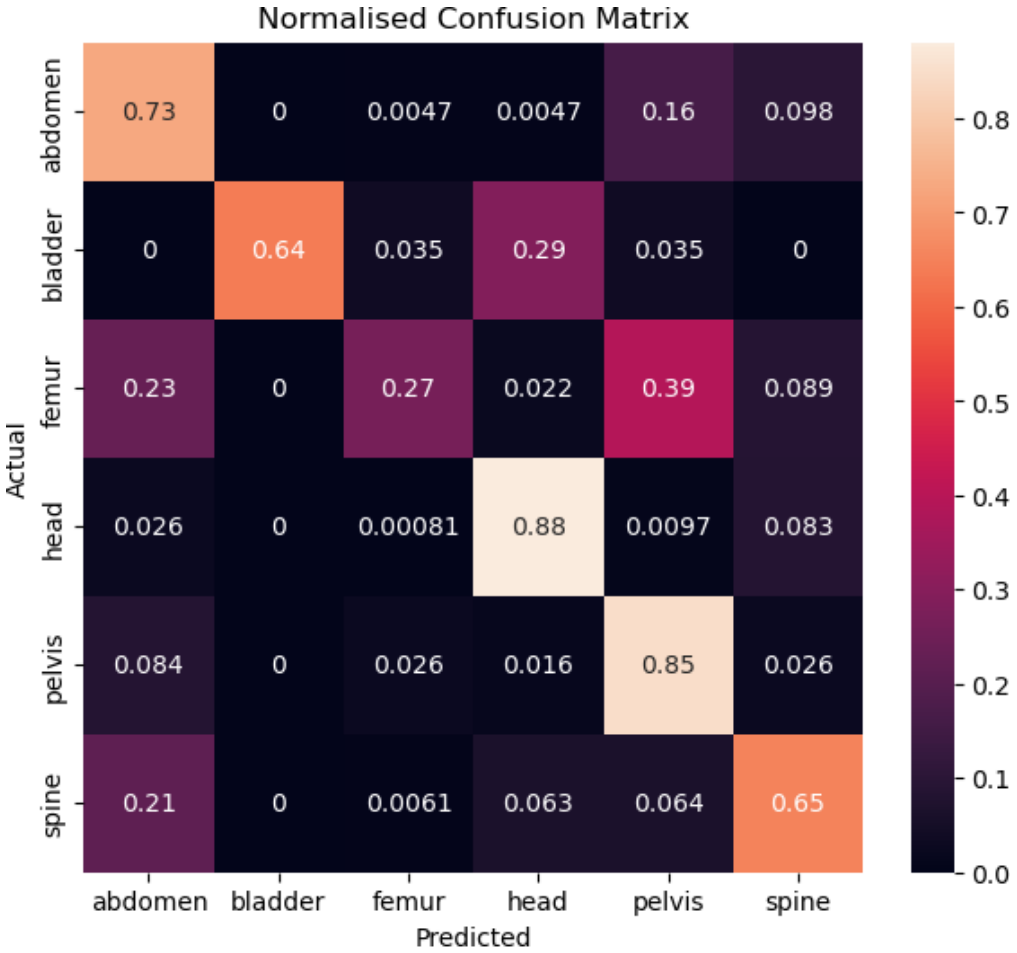

Figure 9: Confusion matrix of the object detector model.

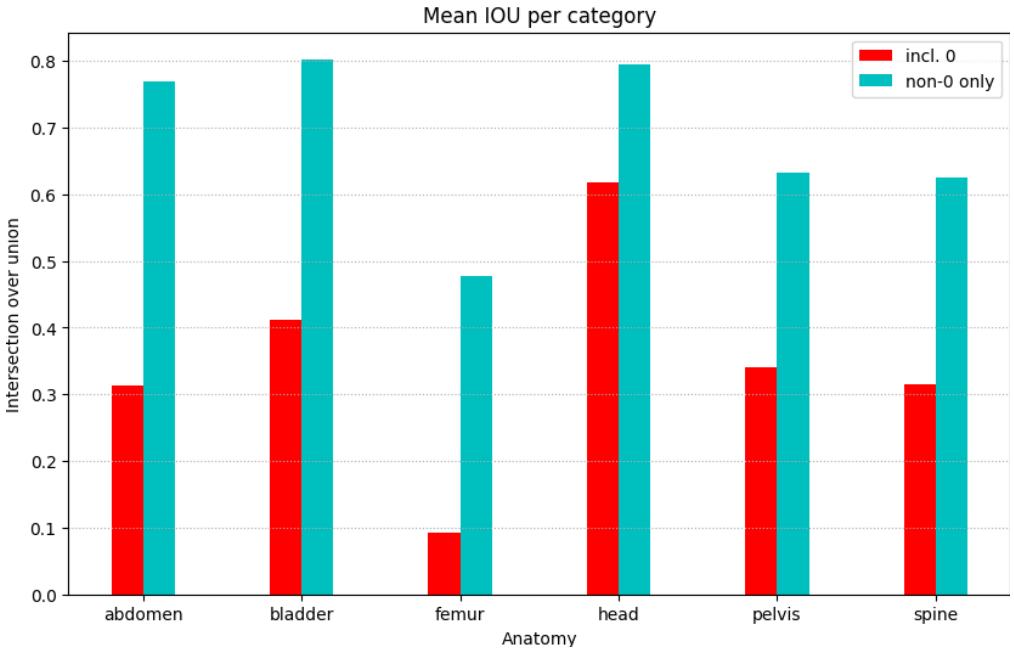

Figure 10: The red bars show the intersection over union of the bounding boxes for each anatomy where the no prediction of a box count as 0 intersection. The cyan bars show only the intersection over union of the boxes that have been predicted.