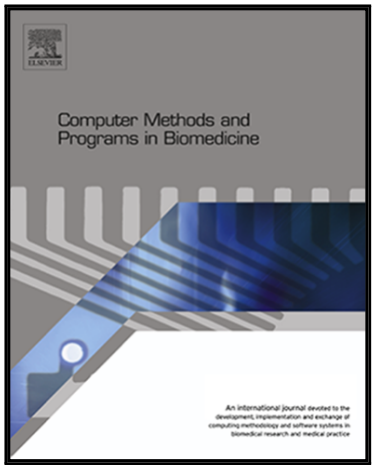
Journal Pre-proof



Public Database for Validation of Follicle Detection Algorithms on 3D Ultrasound Images of Ovaries

Božidar Potočnik, Jurij Munda, Milan Reljič, Ksenija Rakić, Jure Knez, Veljko Vlaisavljević, Gašper Sedej, Boris Cigale, Aleš Holobar, Damjan Zazula

PII: S0169-2607(20)31454-1
DOI: https://doi.org/10.1016/j.cmpb.2020.105621
Reference: COMM 105621

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 6 December 2019
Accepted date: 15 June 2020

Please cite this article as: Božidar Potočnik, Jurij Munda, Milan Reljič, Ksenija Rakić, Jure Knez, Veljko Vlaisavljević, Gašper Sedej, Boris Cigale, Aleš Holobar, Damjan Zazula, Public Database for Validation of Follicle Detection Algorithms on 3D Ultrasound Images of Ovaries, *Computer Methods and Programs in Biomedicine* (2020), doi: https://doi.org/10.1016/j.cmpb.2020.105621

Highlights

- Publishing of the USOVA3D public database of annotated 3D ovarian ultrasound images
- Ovaries and follicles annotated by two gynaecologists
- Design of a verification protocol for unbiased assessment of detection algorithms
- Introduction of two advanced algorithms for follicle and ovary detection
- Inter-rater variability and baseline performance assessed on this database

Public Database for Validation of Follicle Detection Algorithms on 3D Ultrasound Images of Ovaries

Božidar Potočnik[a,*], Jurij Munda[a], Milan Reljič[b], Ksenija Rakić[b], Jure Knez[b], Veljko Vlaisavljević[c], Gašper Sedej[a], Boris Cigale[d], Aleš Holobar[a], Damjan Zazula[a]

[a]*Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia*
[b]*University Medical Centre of Maribor, Slovenia*
[c]*IVF ADRIA Consulting, Maribor, Slovenia*
[d]*LogicData, Maribor, Slovenia*

Abstract

Background and objective: Automated follicle detection in ovarian ultrasound volumes remains a challenging task. An objective comparison of different follicle-detection approaches is only possible when all are tested on the same data. This paper describes the development and structure of the first publicly accessible USOVA3D database of annotated ultrasound volumes with ovarian follicles.

Methods: The ovary and all follicles were annotated in each volume by two medical experts. The USOVA3D database is supplemented by a general verification protocol for unbiased assessment of detection algorithms that can be compared and ranked by scoring according to this protocol. This paper also introduces two baseline automated follicle-detection algorithms, the first based on Directional 3D Wavelet Transform (3D DWT) and the second based on Convolutional Neural Networks (CNN).

Results: The USOVA3D testing data set was used to verify the variability and reliability of follicle annotations. The intra-rater overall score yielded around 83 (out of a maximum of 100), while both baseline algorithms pointed out just a slightly lower performance, with the 3D DWT-based algorithm being better, with an overall score around 78.

Conclusions: On the other hand, the development of the CNN-based algorithm demonstrated that the USOVA3D database contains sufficient data for successful training without overfitting. The inter-rater reliability analysis and the obtained statistical metrics of effectiveness for both baseline algorithms confirmed that the USOVA3D database is a reliable source for developing new automated detection methods.

Keywords: 3D ultrasound images of ovaries, Detection of ovarian follicles, Public database,

1. Introduction

In 1978, the first baby was born after conception by In-Vitro Fertilisation (IVF). The IVF success depends greatly on ultrasound examinations of women's ovaries that are scanned to locate and assess the ovarian follicles best for eggs' retrieval.

Ovarian follicles are pockets of tissue filled with the fluid that protects eggs. After a period of growth inside the ovary, they reach 1 to 2 mm in diameter, which is called the antral stage. In a normal menstrual cycle, the antral follicles grow up to 5 mm in a few days. This qualifies them as dominant follicles essentially ready to ovulate. Their diameters increase by approximately 1 mm a day, and they measure from 18 to 30 mm just before ovulation [1, 2]. Fig. 1 depicts a sample 3D ovarian ultrasound image with follicles annotated by an expert in 2D views of selected cross-sections through the volume (from top-left to bottom-left) and in a 3D view (bottom-right).

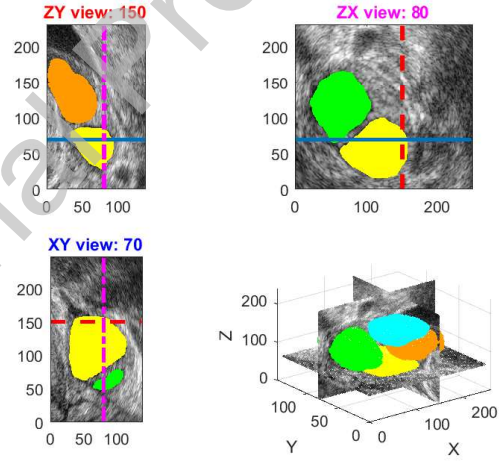


Figure 1: Sample ultrasound volume from the USOVA3D database: a 3D view of selected cross-sections through volume with annotated follicles (bottom-right) and 2D views of annotated follicles superimposed on selected cross-sections (from top-left to bottom-left).

*Corresponding author.
E-mail address: bozidar.potocnik@um.si (B. Potočnik)

A successful IVF depends on two key decisions: Regular observation of follicular development and proper hormone doses. Medical experts carry out ultrasound examinations every 1 to 3 days, which results in sets of ultrasound recordings that have to be analysed. Manual assessment is time consuming and tiresome, as it takes at least 10 minutes after each examination.

Therefore, automated detection of ovarian follicles in ultrasound images is much appreciated when its effectiveness is comparable with experts' annotations. New computer algorithms appear in parallel with the development of improved medical ultrasound devices. Detection of follicles is based on the B-mode ultrasound property that shows follicles as darker regions on a brighter background. All modern ultrasound devices support 3D recording, and contain computer algorithms that help Sonographers recognise the observed 3D structures.

Recently, most attention in clinical practice was drawn by the programme called SonoAVC [3, 4, 5]. It is built in the General Electric ultrasound devices and helps detect ovarian follicles in 3D recordings. Its operation is not fully automated, for it depends on the initial delineation of the ovary, erasure of erroneous follicle detections, and insertion of missing follicles–and all this has to be done manually. Despite that, SonoAVC makes the process of follicle detection much easier, and yet more advanced algorithms should yield higher effectiveness and, in particular, higher autonomy [6].

A better computer detection of follicles still challenges many researchers [6, 7]. They build their own databases of ultrasound recordings, with follicles annotated by medical experts. These segmentations serve as references for a verification of the effectiveness the algorithms they develop achieve. However, such approach generates at least two drawbacks:

- Detection results obtained this way by different computer algorithms cannot be compared unbiased, because they are not based on the same set of ultrasound images and manually segmented references;
- Different metrics are used to assess the effectiveness and accuracy of algorithmically detected follicles in comparison to expert-defined references.

An objective comparison of different follicle-detection approaches is only possible when all are tested on the same database. To the best of our knowledge, such a database is not publicly available. To close this gap, we (i.e., researchers from the Faculty of Electrical Engineering and Computer Science at the University of Maribor, UM FERI, and the IVF experts from the

University Medical Centre of Maribor, UKC), decided to build, and make publicly accessible a database of annotated ultrasound images with ovarian follicles.

This paper describes the development and structure of the USOVA3D database. The ovarian ultrasound volumes were evaluated and annotated independently by two experts. We have also implemented a web service to access and manipulate this database. A general verification protocol for unbiased assessment of algorithms was added to this database in the form of support routines. A combination of the annotated USOVA3D database and verification protocol enables unbiased comparisons of different follicle-detection algorithms and their ranking by performance. One of the goals of this research is to publish tentative values of the follicle and ovary detection performance, measured on testing data from this database. Therefore, we introduced the two baseline algorithms used in this paper. The first procedure, that uses 3D waveform transformations, is based on heuristics or 'hand-crafted' features, while the second is a learning-based procedure grounded on the Convolutional Neural Networks theory.

The contribution of this research work is summarised in

1. The development and publishing of the USOVA3D public database of annotated 3D ultrasound images of ovaries, whereat annotations are provided by two independent experts;
2. Design of a verification protocol for unbiased assessment of detection algorithms;
3. Introduction of two baseline algorithms for follicle and ovary detection.

2. Related work

About three decades ago, initial approaches in ovarian follicle detection used 2D ultrasound images and looked for follicles as darker regions by various algorithms, such as heuristic graph searching, optimal thresholding, 2D region growing, cellular automata, and cellular neural networks [7]. Sensitivity and efficiency (e.g., follicle Contour Mean Absolute Distance, C MAD) of those 2D solutions were up to 80% and with around 1.1 mm C MAD on each 2D image. Recently, two effective 2D methods have been published [8, 9]. Both are based on active contours, texture feature extraction and selection (Particle Swarm Optimisation), and multilayer perceptrons for the follicle classification. Sensitivity of the two methods was near 100%, while specificity was slightly lower, 93% and 96.8%, respectively. However, the authors did not report the distance metrics of detected follicles (e.g., C MAD). Two similar approaches based on CNN and U-Net architecture have also recently emerged for segmenting 2D ovarian ultrasound images [10, 11].

Sensitivity was reported around 80% and specificity 73% [11]. The quality of the detected follicles was assessed by the pixel-wise Dice Similarity Coefficient (DSC), which was around 0.69 [11] and 0.76 [10]. To construct a full 3D representation of the follicles out of a few 2D cross-sections (even not knowing their exact pose), was, and is, very unreliable [12].

Only the development of improved medical ultrasound devices that support 3D recording paved the way to new computer algorithms tackling the 3D ultrasound images as a whole. Some of them extend 2D solutions (e.g., the Kalman filter based approach to ovarian image sequence segmentation [12]), while some incorporate new ideas, such as approaches based either on continuous wavelet transform, level sets, or on inspecting full rotations of the gradient vector by tracing the follicle boundary in 3D [7].

Recently, the most efficient follicle detection methods in 3D ultrasound images of human ovaries have been reported by Chen et al. [13], by our team based on Directional 3D Wavelet Transform (3D DWT) [6], and by the SonoAVC proprietary solution incorporated in the General Electric ultrasound devices for automatic volume calculation [4, 5]. Chen et al. trained a probabilistic framework on annotated ovarian volumes and built models of ovaries and follicles. The sensitivity of this approach was around 78%, while the Volume Match (V Match) of detected follicles with an expert's annotations was around 80%. The mean Euclidean distance between surfaces of detected and annotated follicles (S MAD) was around 1.87 mm. The 3D DWT-based approach from our previous work [6] demonstrated 88% sensitivity and 85% specificity. Volume match was around 98%, while the S MAD was 0.31 mm. It should be stressed that this paper upgrades this method. The SonoAVC is a semi-automated detection approach that is based on manual outlining of the ovarian region by proceeding through the entire ultrasound volume in preselected rotations. Afterwards, SonoAVC interpolates the annotations throughout the volume. This solution could miss follicular regions, it can under- or overestimate the regions, and it can merge two or more regions. Without further manual post-processing (e.g., deleting False Positives, cutting the merged regions), the results may not be acceptable [6, 7]. Nevertheless, several comparison studies have confirmed a fairly accurate estimation of follicular volumes by SonoAVC [6]. Reported results without the post-processing step are as follows: Sensitivity around 52%, with Follicle Misidentification Rate (FMR) low at 9%, while the V Match was around 90% [7].

In Table 1, we gathered some important results as published by recent follicle detection

methods. It should be emphasised that these methods were validated on private data only, which prevents a direct comparison of their results, and that the ovarian images/volumes were typically annotated by a single expert. We discuss the reported private datasets in the next subsection. As a matter of fact, no publicly available annotated database on ovarian (2D or 3D) ultrasound images with follicles has been created so far.

Table 1

2.1. Reported datasets on annotated ovarian follicles

The entire development and testing of 2D ultrasound follicles detection methods were based mainly on private data [7]. It also applies to recent work in this research area. Marques [11] employed 99 B-mode transvaginal ovarian ultrasound images, with image dimensions ranging from 192×200 to 192×620 pixels. Single annotation was available for ovary and follicles for each image. A similar approach is reported in [10]: 87 B-mode images, a private dataset, and a single medical expert rater. The images were acquired by using the Ultrasonix SonixTouch Q+ device. A different approach was undertaken in [9], where 60 ovarian ultrasound images were picked up from the publicly available websites. Afterwards, these images were annotated by a Gynaecologist. In [8], the images were obtained from the US National Library of Medicine (National Institute of Health), but the authors did not provide any details on the image size, annotation procedure, gold standard, etc.

The situation is much the same in the field of 3D ovarian ultrasound images. Early work was based either on private freehand 3D human ultrasound data or on private bovine ovaries datasets. These approaches were, besides on animals, also tested on a small number of humans (11 and 6 patients, respectively), whereat practically no information about these volumes were given [7]. Among the more important newer methods, the method in [13] employed 501 human ovarian ultrasound volumes, namely 400 for training and 101 for testing. The median volume size was $243 \times 177 \times 121$ voxels with voxel size around 0.8 mm. This private dataset was accompanied by a single annotation for ovary and follicles. Proprietary solution SonoAVC was validated thoroughly on 31 ultrasound volumes obtained from 14 patients [5]. This private data set was acquired by using the Voluson 730 utrasound device. A single examiner evaluated the volumes and provided ovary/follicles' annotations. Our 3D DWT-based solution [6] was also validated on a private dataset; 30 volumes were acquired by the Voluson 730 and the Medison Accuvix XQ ultrasound devices. Mean volume size was $165 \times 141 \times 181$ voxels with 0.2 mm voxel size. Two

medical experts provided reference annotations for ovaries and follicles. It should be noted that this dataset forms the core of the USOVA3D database.

Characteristics of the abovementioned datasets are summarised in Table 1.

3. A database of annotated 3D ultrasound images of ovaries–USOVA3D

A database of 3D ultrasound images of ovaries, USOVA3D, was constructed by a team of Gynaecologists and Sonographers from UKC and researchers from UM FERI. Medical experts rated the images, and provided manual segmentations of ovaries and follicles. The database structure, web tools, and annotation protocols were developed at UM FERI. We have followed the designs of established publicly accessible databases from various research fields, such as [14, 15, 16].

All women participating in this research signed informed Consent, and their anonymised ultrasound recordings were added to this database. The Ethical Committee at UKC approved the research activities and the database construction.

This database includes 35 segmented volumes of women's ovaries. Details about the USOVA3D database are gathered in Table 2. Each USOVA3D entry consists of one 3D image and the corresponding segmentations of the ovary and follicles. Manual segmentations were contributed by two independent Sonographers from UKC for each image. They used the ITK-SNAP [17] tool, which imports and exports files based on the VTK data format [18]. All the segmentations were, therefore, implemented in 3D, and each saved in a separate VTK-formatted file. Files with ovarian segments describe the ovarian region by the voxel value 1, and the background by 0, whereas the files with follicular segments denote separate follicles by consecutive numbers beginning from 1, and the background by 0. The sizes of the two segmentation volumes equal the size of the corresponding ultrasound image.

The database entries link five files, each as follows: A file with a 3D image, two files with ovarian and two files with follicular segmentations, produced by two independent raters. All the data files implement the VTK data format.

All USOVA3D database entries are separated into a training and a testing set. The training set consists of 16 entries, whereat each single entry contains the original 3D image and annotations of the ovary and follicles of both raters. On the other hand, the testing set consists of 19 entries, whereat just the original 3D images are available. Such splitting and design of the database with

no 'ground truth' available for the testing set enables the evaluation of recognition algorithms by using our USOVA3D web services only. This leads to a more fair validation and comparison of algorithms.

Table 2

4. USOVA3D web portal and services

The USOVA3D database is accessible at the web address https://usova3d.um.si/. Primarily, this web portal provides access to the training and testing sets. Utility routines for manipulating database entries (e.g., loading and storing data in the VTK format) are added as well. Auxiliary routines that implement our evaluation protocol are also available. All routines are written in the Matlab scripting language.

This web portal includes an option to upload segmentation results obtained on the USOVA3D testing set by an external algorithm. Web services for validating the segmentation results are triggered by uploading. The obtained report is, afterwards, inspected manually by the UM FERI personnel. Finally, a validation summary is published for the algorithm on our web portal in a joint Table, whereat a position is determined by the algorithm's performance and expressed by the overall score. Such ranking of the results implies a direct comparison of the performance of different detection methods.

We developed the USOVA3D web portal independent of the operating system and server architecture and, for public access on our servers, it runs in the Linux environment. A combination of Apache server [19], SQL database and WordPress content management system [20] is used. A modified image processing system running in a cloud is used for the validation [21, 22].

4.1. Access to the database

Only authorised users are able to access the database. Typically, users sign in to this web portal by using the Google Sign-In service [23], although a local password can be requested as well. Because we are dealing with medical data that has been collected in research projects, users must first accept a Research License. As there is a potential for misuse and non-legitimate access to the database, we keep information about logins and downloads, which requires the user's consent in accordance with the General Data Protection Regulation [24].

Users see our database split into two parts. The first represents a training set that provides access to all five files for one database entry, namely to a 3D image, to the two files with ovarian annotations, and to the two files with follicle annotations (one file per rater). The second part of the database contains a testing set, of which users are only able to access 3D images. Annotations for the testing set serve just to validate the segmentation results that users send for verification.

4.2. Validation of submitted follicle detections and reporting

The USOVA3D portal allows validation on our site only. Users have to upload their results in the prescribed format. Information about the detection method (e.g., algorithm name, reference) and its performance, are collected in a Table. Measured metrics are published in a condensed form. Algorithms are ranked according to the overall algorithm's score.

5. Verification and validation protocol

According to recent results, medical experts perform better in estimating the ultrasound-image structures than any algorithm so far [1, 3]. The absolute truth cannot be reached by ultrasound observations and without invasive biopsies, which is not available in our case. For this reason, manual annotations are believed to be the closest approximations of ground truth; by combining estimates of several independent raters, the statistical significance increases and inconsistency decreases. Every volume in our database is, therefore, accompanied by two independent segmentations of the ovary and follicles. The two experienced Sonographers may have decided for a different number of follicles within the same volume, and different shapes of the follicles. The higher their agreement, the greater is the confidence in their segmentations. We reveal here the quantitative measures that we built into our web services to establish inter-rater agreement, and to verify any computer algorithm against the referential raters' segmentations.

There is a minor limitation concerning the statistical indices, due to missing True Negatives; namely, it is not possible to find out which of the image structures the raters could have declared follicles, but they didn't. Hence, our statistics depend on the available annotated regions of two raters that are taken as references. These references are then compared to either a non-referential rater, or the algorithm under inspection. In any comparison, we call the compared regions the referential and the detected ones. Our estimation approach is as follows.

When a detected region covers one referential region at least in the smallest extent, the count of True Positives is incremented (TP). Indeed, this arbitrary decision generates sensitivities and precisions higher than with any other threshold of region overlapping, but does not corrupt the comparison results in Eq. (5), which is the main goal of the USOVA3D database. When a detected region doesn't correspond to any referential region, the count of False Positives is incremented (FP), and when a referential region has no contact with any of the detected regions, the count of False Negatives is incremented (FN). In the case of several referential regions covered by one detected region only, the biggest referential region increments the TP count, and the remaining referential regions the FN count. Finally, if several detected regions cover one referential region only, the biggest one increments the TP count, while all the others increase the FP count.

Sensitivity, S, and precision, P, are calculated as follows:

$$S = \frac{TP}{TP + FN}, \qquad P = \frac{TP}{TP + FP}. \tag{1}$$

Besides being able to detect follicles, it is clinically important to detect their boundaries as accurately as possible. For this reason, we have introduced product $\rho_1\rho_2$ [25] and we extend it for 3D here. Value ρ_1 equals the ratio of the intersecting volume of detected and referential follicular 3D regions and the volume of the referential follicle itself, whereas ρ_2 equals the ratio between the intersecting volume and the detected follicle volume. The closer their product is to 1, the better is the matching of the detected and referential follicles. For example, $\rho_1\rho_2 = 0.65$ corresponds to an 80% overlap.

Here, we introduce thee additional measures:

1. The ratio of the total volume of correctly detected follicles (V_d) and the total volume of all the referential follicles (V_r):

$$r_V = \frac{V_d}{V_r}; \tag{2}$$

2. The mean Euclidean distance (in voxels) between the surfaces of correctly detected and referential follicles (i.e., metric S MAD):

$$\overline{e}_{fol} = \frac{1}{M} \sum_{i \in (\text{all superficial voxels})} E\left(\mathbf{p_d}(i) - \mathbf{p_r}(i)\right), \tag{3}$$

where M stands for the total number of superficial voxels on the detected follicle, E for the Euclidean distance operator, $\mathbf{p_d}(i)$ for coordinates of the i-th detected voxel, and $\mathbf{p_r}(i)$

the coordinates of the i-th referential voxel. We determine the superficial voxels on follicular surfaces by subtracting the morphologically eroded region from the entire region of a processed follicle. Then, we calculate spatial distances from each voxel on the detected follicle surface to the nearest superficial voxels on the corresponding referential follicle. The shortest spatial distance found enters Eq. (3);

3. The mean absolute difference (in voxels) between the diameters of equivalent spheres that have the same volumes as the detected and referential follicles:

$$\overline{d} = \frac{1}{N} \sum_{i \in (\text{all follicles})} |d_d(i) - d_r(i)|, \tag{4}$$

where N stands for the number of all the corresponding follicles, $d_d(i)$ for the equivalent diameter of the i-th detected, and $d_r(i)$ for the equivalent diameter of the i-th referential follicle, respectively.

5.1. Combining several statistical measures

The statistical measures presented above are calculated for the verified algorithm for each 3D volume. To compare different detection algorithms, we can, of course, sort them by performance according to the criteria chosen. Often, the best algorithm is sought according to all given criteria. In our protocol the so-called combined score is computed. It should be emphasised that the metrics S, P, r_V and $\overline{d}$ are calculated for the 3D volume as specified in Eqs. (1), (2) and (4), while the product $\rho_1\rho_2$ and $\overline{e}_{fol}$ from Eq. (3) are calculated for each follicle individually. Therefore, the means of $\rho_1\rho_2$ and $\overline{e}_{fol}$ are computed over all detected follicles denoted as $\overline{\rho_1\rho_2}$ and $\overline{e}$ and used in the combined score calculation. The combined score, ξ, for the selected algorithm is calculated for each 3D ultrasound volume as a linear combination of statistical metrics as:

$$\xi = 20\left(S \cdot P + \overline{\rho_1\rho_2} + r_V + sc(\overline{e}) + sc(\overline{d})\right), \tag{5}$$

where the quasi normalisation function sc is defined as:

$$sc(x) = \begin{cases} 1 - 0.1 \cdot |x|, & \text{if } |x| \leq 10 \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

The constant in (6) was determined empirically based on the spatial resolution of 3D volumes in our database (10 voxels corresponds to 2 mm; a 2 mm error in distance is not acceptable). It applies: The higher the combined score ξ, the better the detection algorithm performs on the

selected volume. The maximal combined score 100 is obtained by a perfect match between the detected and referential follicles (annotated by a single rater).

5.2. Integrating the combined scores of several independent raters

When dealing with several raters, we obtain one combined score ξ for the volume by each rater. We must in some way integrate these combined scores in order to determine the performance of a detection algorithm by considering all raters. It is quite possible that some annotations of independent raters disagree. To assess the efficiency of detection algorithms, either only those annotations with full raters' agreement can be considered, or the efficiency is determined for each rater separately and their arithmetic mean is calculated afterwards. Our assessment protocol assigns higher importance to annotations where all raters agree, but we do not want to discard or disregard non consensual annotations that are taken into account with a lower importance. We give an explanation for two independent raters in the sequel, but the idea can be generalised easily. All reference follicles of both raters are divided into two groups: i) Those marked by both raters and ii) Those not marked by both raters. Statistical metrics calculated for the referential follicles in group i are then taken into account in the final score, with a greater weight than the metrics for the referential follicles in group ii.

First, two new virtual raters, rater vr_1 and rater vr_2, are created from raters r_1 and r_2. The virtual rater vr_1 is obtained by retaining only those referential follicles of rater r_1 that were also annotated by rater r_2. Similarly, a virtual rater vr_2 is created by retaining only those referential follicles of rater r_2 that were also annotated by rater r_1. Four different combined scores are, thus, obtained for the selected volume, and the detection algorithm, depending on the referential follicles of which rater are used in the validation (i.e., combined scores ξ_{r_1}, ξ_{r_2}, ξ_{vr_1}, and ξ_{vr_2}). It should be emphasised that metrics for the referential follicles in group ii are part of combined scores ξ_{r_1} and ξ_{r_2}. Finally, these scores are integrated into the final score ξ_{vol} of the algorithm performance on the selected volume by a linear combination, whereat the consensually annotated referential follicles (group i) are taken into account with two times the weight of an individual rater's score. Calculation of the final score ξ_{vol} is, therefore, implemented as:

$$\xi_{vol} = \frac{1}{6}\left(\xi_{r_1} + \xi_{r_2}\right) + \frac{1}{3}\left(\xi_{vr_1} + \xi_{vr_2}\right). \tag{7}$$

5.3. The overall score

We obtained the algorithm's detection effectiveness for only one database entry by applying the above procedure. In order to assess the effectiveness on the entire database, we need to summarise these scores over all entries. The fact is that, whatever statistics are used (e.g., mean, trimean or median value), some ambiguity will always be introduced in the final assessment. Namely, selected statistics will emphasise just certain properties of results, and will neglect the rest (e.g., mean statistics consider both tails of the result distribution very poorly). Our aim was to deploy statistics that rely on the entire distribution of results as much as possible. Therefore, we implemented this compromise solution. If $\xi_{vol,i}$ denotes the final score for the i-th entry, final scores for all entries in the database distribute to five classes, C_k. An average is calculated for each class C_k as:

$$\mu_{C_k} = \mathbb{E}(\xi_{vol,i}); \quad \xi_{vol,i} \in C_k, \tag{8}$$

where $\mathbb{E}$ denotes a mathematical expectation operator, and classes C_k correspond to five equally distributed percentile boundaries with respect to all final scores (minimum to the 20th percentile, 20th to the 40th percentile, etc.).

An overall algorithm score, ξ_{alg}, is finally determined as the weighted sum of calculated means for the classes C_k, whereat classes in the middle of the resulting distribution are weighted heavier than those in the tails. This score is calculated as:

$$\xi_{alg} = \frac{1}{9}(\mu_{C_1} + \mu_{C_5}) + \frac{2}{9}(\mu_{C_2} + \mu_{C_4}) + \frac{1}{3}\mu_{C_3}. \tag{9}$$

Eq. (9) calculates the algorithms' performance with respect to the detection results on the USOVA3D training set. This value also ranks the algorithms. Let us emphasise that a perfect detection algorithm would have ξ_{alg} equal to 100. It should also be stressed that Eq. (9) does not minimise error representation, but combines results with less smoothing. Number of classes $k = 5$ was preselected, and could also be increased for bigger databases, whereat Eq. (9) needs to be modified in such cases.

6. Two examples of automated ovarian follicle detection

The USOVA3D database is expected to be a reliable source of segmented regions that two experts recognised as follicles. This is taken as the best 3D approximation of actual ovarian structures, and, thus, considered a ground truth for subsequent verification. To estimate its quality, we

want to answer two important questions: (a) Can we trust the annotations provided by the two Sonographers, and (b) How useful is this database when developing new automated detections of follicles?

The answer to the first question is based on the consistency of annotations that can be measured by inter-rater reliability (see Section 7.1). To answer the second question, we developed and evaluated two different follicle detection procedures. The one which is based on Convolutional Neural Networks (CNN) [26] uses both the training (for the detection model development) and testing (for the detection efficacy assessment) sets of this database, while the other one, which applies Directional 3D DWT [6], needs only the testing set.

6.1. CNN-based detection method

Although we are dealing with volumes, we propose naive solutions based on a 2D Convolution Neural Network for a 'baseline' follicle and ovary detection procedure. Three-dimensional CNNs and/or solutions based on Recurrent Neural Networks represent directions for future research.

Our solutions are based on the well-known U-Net architecture [27], which was introduced for bi-level segmentation of biomedical images. U-Net CNN expects a 256×256 pixels colour image as input, and returns a 2D binary segmented image of the same size. We left the U-Net architecture practically unchanged; only the input layer was adapted so that 2D grey-scale images can be inserted into the net. The core of our work was in the preparation of training data, network training, and combining partial segmentation results.

Let us first explain the follicle detection procedure. We employ volumes and corresponding reference follicles from the learning set for the training. Annotations of two raters are used for each volume. The inputs to the U-Net are 2D images, therefore, the 3D ultrasound volumes and, of course, the reference follicles, were 'cut' into planar cross-sections. To capture at least some spatial correlation between voxels in volume, the volume is 'cut' with respect to all three directions (planes), giving the planar cross-sections in the x, y, and z directions. By using three orthogonal cross-sections in CNN rather than one 3D volume sliced into cross-sections in a single direction and then fed into CNN, we avoid the problems identified in the 2D follicle detection methods based on U-Net [10, 11], and also increase the size of the training set significantly (i.e., training data are in some way augmented). All cross-sections are scaled to the size of 256×256. The aspect ratio of images is not maintained in this naive solution. 8297 cross-sections, or

$2 \cdot 8297 = 16594$ data contributed by two raters, are obtained in this way for 16 ultrasound volumes from the learning set. These data were then split into a training set and a validation set in a 4:1 ratio. Each annotated cross-section from the training set is used individually as one sample during training.

Our network was trained by the Adam optimisation algorithm [28]. An early stopping, and strategy of decreasing the learning rate when reached the plateau of loss function, were applied by training [28]. Hyper-parameters, such as Learning Rate (LR=0.01), LR decay rate (0.5), number of steps for early stopping (50) and plateau patience (10) were determined by a grid-search. Training data were also augmented (images were translated, flipped horizontally and vertically, and corrected with respect to brightness). The loss function L was defined as the sum of binary cross-entropy loss (L_{Ent}), loss based on Dice Similarity Coefficient (L_{DSC}), and loss based on the $\rho_1\rho_2$ product ($L_{\rho_1\rho_2}$, see also Section 5) as:

$$\begin{aligned} L(y,\hat{y}) &= L_{Ent}(y,\hat{y}) + L_{DSC}(y,\hat{y}) + L_{\rho_1\rho_2}(y,\hat{y}) = \\ &= -\frac{1}{Num}\sum_i y_i \log \hat{y}_i + \left(1 - \frac{2\sum_i y_i\hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i}\right) + \left(1 - \frac{(\sum_i y_i\hat{y}_i)^2}{\sum_i y_i \sum_i \hat{y}_i}\right), \end{aligned} \tag{10}$$

where y_i is true (annotated) and $\hat{y}_i$ is the predicted value of the i-th pixel, while Num denotes the number of all pixels in the cross-section. It was determined empirically that all terms contribute to the loss L equally. It should be emphasised that all losses were calculated at the level of same-laying pixels from the segmented and reference 2D image, and not at the follicular level. After training, the U-Net weights were set to weights of the epoch in which the lowest loss L was obtained on the validation set.

The trained CNN was applied to detect follicles in volumes, as shown in Fig. 2. The volume was first 'cut' in all three spatial directions into planar cross-sections. After scaling, the cross-sections were inserted into the network, and a segmented 3D volume was constructed from the segmented 2D slices. This volume was scaled appropriately to its original size. We had cross-sections in three directions through the volume, consequently, we got three segmented volumes after this processing step. The three segmented volumes were then combined into the final segmentation result, using the majority vote procedure. This result was processed further by 3D morphological operators, due to minor computational errors and rounding-off errors throughout the entire processing pipeline.

The above described approach was used in exactly the same way (including hyper-parameters)

to develop the 'baseline' method for ovary detection. The only difference was that the CNN was trained with annotated ovaries instead of annotated follicles.

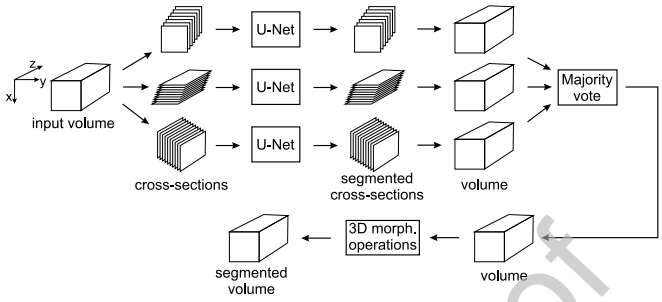


Figure 2: Schematic depiction of ovary/follicle detection in 3D ultrasound volumes using a CNN-based procedure.

6.2. Method based on Directional 3D Wavelet Transform

We showed in [6] that the wavelet transform improves detections in 3D images of ovaries. The voxel grey levels can be considered random variables, whose distribution fits Gaussian mixtures with up to four peaks. 3D DWT, as introduced in [6], increases distances among the Gaussian components, while their deviations do not change. Thus, a separation between wavelet-transform coefficients that belong to the follicles and the coefficients that surround follicles increases and becomes more reliable. The separation threshold is defined by the maximum distance of the means of Gaussian components [6], and is proportional to the wavelet-transform scale.

This procedure detects candidate follicular regions. Their number can be greater than the number of follicles, which is due to other ovarian tissues, artefacts, noise, etc. Recognition of true follicles is, therefore, a challenging task. A solution to this problem was proposed in [6], based on the properties of the individual follicles only, without any knowledge about the ovary. Should the extent of the ovary be known, this was important information to discard all candidates outside the ovary. On the other hand, the true follicles themselves outline the space of an ovary.

Bearing this in mind, we propose a novel follicle detection approach in the following eight steps:

1. After applying the procedure from [6], we compute the means and standard deviations of grey levels within the candidate follicular regions, and also for 1-mm shells surrounding

the regions: $\mu_{f,i}$, $\sigma_{f,i}$ and $\mu_{o,i}, \sigma_{o,i}$ for the i-th follicle and its close surroundings, respectively.

2. For each candidate follicle that verifies the condition

$$\mu_{f,i} - \mu_{o,i} > \frac{\sigma_{f,i} + \sigma_{o,i}}{2}, \tag{11}$$

two histograms are constructed, one for grey levels within the follicle, $h_{f,i}(n)$, and one for grey levels in 1-mm shells outside the follicles, $h_{o,i}(n)$. The two histograms must cover the same span of grey levels, i.e. between 0 and 1, as they appear in the context of the approach in [6]. Indices of candidate follicles that verify the condition (11) form the set $\mathbf{I_c}$.

3. A threshold, η_i, is defined for the i-th candidate; $i \in \mathbf{I_c}$:

$$\eta_i = \text{maxarg}_x \left(\sum_{n=0}^{x} h_{o,i}(n) \sum_{m=x}^{N_h-1} h_{f,i}(m) \right) \tag{12}$$

where N_h stands for the number of histogram bins.

4. Threshold η_i determines the measure, ϑ_i, of grey-level unlikeness, which is important for discerning the follicles from their background:

$$\vartheta_i = \frac{\sum_{n=\eta_i}^{N_h-1} \left(h_{f,i}(n) - h_{o,i}(n) \right)}{\sum_{n=\eta_i}^{N_h-1} h_{f,i}(n)}. \tag{13}$$

5. We concentrate on dominant follicles close to the centroid of all candidate regions (i.e., 'ovarian centroid'). Dominant follicles are considered those having the diameter of a volumetrically equal sphere greater than 5 mm, and a displacement of their centroids from the ovarian centroid less than 30% of the observed 3D image diagonal. After collecting indices of selected dominant follicles in set $\mathbf{I_d}$, a threshold ϑ_{min} is obtained as:

$$\vartheta_{min} = \max_i \left([\vartheta_i, 0.52] \right); i \in \mathbf{I_d}. \tag{14}$$

6. An approximation of the ovarian region is now outlined by a convex hull around the follicles with $\vartheta_i > \vartheta_{min}; i \in \mathbf{I_d}$.

7. The approximated ovary filters out most probable follicles. The i-th candidate is assigned to the group of recognised follicles if $i \in \mathbf{I_c}$, $\vartheta_i > 0.5\vartheta_{min}$, and 45% of its volume lies inside the approximated ovary.

8. A refined and final approximation of an ovary is set up by a convex hull containing all the recognised follicles.

The numerical parameters in steps 5 and 7 need more explanation. The fact that the dominant follicles exceed 5 mm in diameter is well known [29]. Why the constant 0.52 in (14)? If two normalised Gaussian distributions with the same standard deviation, σ, intersect at $\mu_1 + 0.5\sigma$ and $\mu_2 - 0.5\sigma$, where μ_1 and μ_2; $\mu_2 > \mu_1$ stand for the means of the two distributions, then the difference in two probabilities computed for grey levels from the intersection point on is 0.52. This theoretical value is rather strict, and many real follicles exhibit a lower one due to noisy Gaussian mixtures. Nevertheless, in our search for dominant follicles, we need to induce this strict lower bound, which is in line with the condition (11).

We also assume that dominant follicles appear close to the centroids of ultrasound images, which is due to the fact that Sonographers focus on the ovary during medical examinations. In general, 3D ultrasound images manifest as cuboids. Suppose that a 3D image is a cube. Its diagonal equals $\sqrt{3}a$, where a is the side length. Taking 30% of this length gives about $0.5a$. In step 5, this stands for the maximum distance between the centroid of the observed follicle and the centroid of all candidates. Actually, the limitation prevents regions at the image boundaries from entering our search for dominant follicles. Indeed, if the volume is cuboidal with rather different side lengths, 30% of its diagonal yields more than a half of the shortest side, but not the other two sides.

The given assumptions turn out to be appropriate, and guarantee detection of the majority of dominant follicles. When these are involved in an approximation of the ovary, it opens a way to less prominent follicles being detected if found within, or close to, the outlined ovary. Detection criteria must only be relaxed a bit for this reason. This decision seems almost arbitrary, see step 7. We decided to halve ϑ_{min} and recognised all corresponding candidate follicles if at least 45% of their volumes lay within the ovary. A simplification will help us explain the last figure. Suppose all follicles and the ovary are spheres. We also assume that a region may be counted as a follicle if its centroid touches the approximated ovarian border, or is inside the ovarian region. In the case of only touching the border, at least 45% of follicular volume lies inside the ovary. The fraction depends on the follicle size, and is bigger with smaller follicles, but never exceeds 50%.

7. Results

We will first examine how well the USOVA3D database is annotated. Therefore, we will evaluate and analyse the variability and reliability of the two raters. Subsequently, we will present the results for both proposed baseline follicle detection approaches on the USOVA3D database testing set.

7.1. Inter-rater variability and reliability

A total of 109 ovarian follicles were identified by each of the two raters in 19 volumes from the USOVA3D database testing set. Rater 1 annotated 5.7 ± 4.5 follicles, while rater 2 denoted 5.7 ± 4.8 follicles on average per volume. Both raters fully agreed on the number and position of the follicles at only 5 volumes. In total, the raters agreed on the denotations for 89 follicles over the entire testing set. The latter means that it can be said with greater certainty that the volumes contain 4.7 ± 3.7 follicles on average. Below, we assessed how much raters agreed on the annotations. We measured their agreement by cross-validating the segmented follicular regions by the metrics explained in Section 5. We took the segmentations of the first rater as ground truth, and estimated the success of the second rater, and then reversed their roles.

First, the combined scores ξ for all testing volumes were calculated using (5). The annotations of rater 2 were treated as a result of the 'detection algorithm', while the annotations of rater 1 were taken into account as ground truth. Since we compared annotations of two raters with each other, it was not possible to calculate the final score ξ_{vol} according to (7), but we treated the calculated combined score value as an approximation of the final score for each volume (i.e., $\xi_{vol} = \xi_{r_2}$ for rater 1). The median over all 19 final scores was 84.5 (min = 70.4 and max = 91.5). The overall score, ξ_{rater}, for rater 2 with respect to rater 1's annotations, calculated by using (9), was 83.1. The roles of the raters were switched afterwards. In this case, the median was 84.1 (min = 72.7 and max = 91.2), and the overall rater 1's score with respect to rater 2's annotations was 83.9. All calculated statistics are gathered in Table 3.

Table 3

We also evaluated how well both raters agreed on ovarian annotations. Naturally, the role of follicles in (5) to (9) was replaced by ovaries when calculating these metrics. The obtained statistical metrics are collected in Table 4.

Table 4

7.2. Quantitative results of the proposed detection methods

Both proposed 'baseline' procedures were applied on the USOVA3D database testing set. The quality and performance of follicle detection were evaluated by using our protocol. Aggregated results over all 19 test volumes are summarised in Table 5. The 3D DWT-based method, with an overall score of 78.2, proved to be more successful than the CNN-based detection procedure that had this rating equal to 72.5.

Table 5

In one of its steps, the 3D DWT-based detection procedure also estimated the ovary in ultrasound volume. Hence, we were able to assess how well the ovary was approximated. The verification protocol described above was applied, except that we replaced the role of follicles with ovaries. The obtained statistical metrics are presented in Table 6.

Table 6

8. Discussion

The results of inter-rater variability and reliability prove high inter-rater agreement, and infer reliability and accuracy of the referential follicle segmentations. The former is underpinned by the fact that the raters agreed almost entirely (more than 82%) that the annotated regions were follicles, except for some non-dominant follicles. The latter are quantified by a high overall (algorithm/rater) score above 83, whereat the $\rho_1\rho_2$ metric averaging over 0.75 (i.e., around 86% overlap between denotations of raters) and the mean absolute difference between the diameters of equivalent spheres that have the same volumes as the follicles annotated by rater 1 and rater 2, is below 0.3 mm, on average.

Our database is available to researchers for testing their algorithms. To show what recent, up-to-date automated detections can achieve, we ran two of our own solutions. Some novelties in our solutions need additional explanations.

A solution based on the U-Net architecture was introduced to verify (and demonstrate) whether our database could be used for developing successful learning-based detection procedures. The graph in Fig. 3 depicts how the loss L varied during training of our follicle detection. It can be seen clearly that, by using the USOVA3D database, it is possible to train the detection algorithm quite successfully and without any data overfitting. A convergence was reached in a reasonable number of training steps. This demonstrates experimentally that the introduced

database, although not extremely large in the number of ultrasound volumes in its initial version, also contains sufficient data for the development of learning-based and self-adaptive detection algorithms.

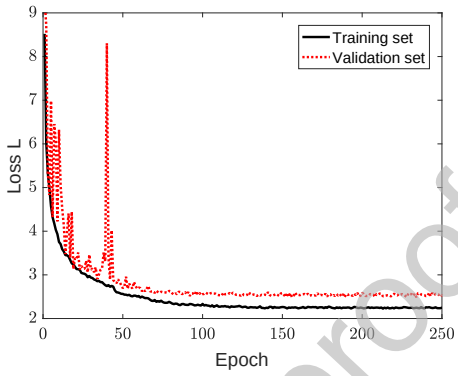


Figure 3: CNN-based follicle detection procedure: Changing of loss L during training with respect to epochs.

We upgraded the 3D DWT follicle detection procedure from [6] with an indirect approach to the detection of ovaries, as derived in Subsection 6.2. For easier comprehension, we are going to comment on some of the algorithmic steps.

We observe each candidate follicle on its local background. It is clear that grey-level distributions within the follicular region and in the shell around the region partly overlap. Empirically, we determined that all the real follicles fulfil the condition (11). This condition eliminates a great deal of erroneous candidate follicles.

For all the remaining candidates, it is important to decide whether the grey-level distribution within the region prevails against those outside the region or not. We propose the solution in steps 2, 3, and 4. Each candidate is assigned a value, ϑ_i, which is 'normalised' in a way according to its local condition. Independent of the changes in local grey-level distributions, these measures are consistently comparable, and can sort out the regions with higher probability of depicting the real follicles.

We have supplemented the USOVA3D database with a carefully established protocol for evaluating results. We consider all raters in this protocol, while a greater emphasis is placed on those annotations where all raters agreed. Annotations with no agreement between raters are not discarded, but are considered with a lower weight. In this way, we have established a

general mechanism for a fairly objective and unbiased assessment of detection results. Various characteristics of the detection procedure are, thus, captured in the overall algorithm score, and, at the same time, algorithms can be ranked with respect to performance based on this score.

Our two baseline procedures were evaluated on the USOVA3D database using the established verification protocol. The algorithm based on the 3D DWT proved to detect follicles with more success. It obtained the overall score of about 5 to 6 (from a maximum of 100) lower than the score calculated within the inter-rater variability analysis. Comparing Tables 3 and 5, the distributions of the algorithm's and raters' detection errors differ on the entire scale. While score maxima (the best detections) coincide, the medians are about 5, and minima about 10 in favour of the raters. Slightly worse performance was achieved by the CNN-based procedure (discrepancy with raters of about 11, with score maxima aligned, medians about 9, and minima about 26 lower, respectively).

Greater mismatch with the raters was measured for the 3D DWT-based algorithm when detecting ovaries. The overall score of our algorithm is about 13 lower compared to the scores obtained with inter-rater variability. The highest portion of error appears below the median score; differences at score maxima and medians are both about 9, whereas the difference at minima increases to about 30. This can be attributed to the way our algorithm approximates the ovary by a convex hull around the most probable candidate follicles. If a follicle is not detected, that part of the ovary is also missed this way. On the other hand, our CNN-based detection procedure estimates ovaries only slightly worse than the raters. Namely, the overall score of our CNN-based approach is just about 4 lower than the inter-observer variability.

9. Conclusion

The main intention of this paper was to introduce a public database of 3D ultrasound images of annotated ovaries. We revealed the database structure and web services that allow users to interact with the database. Part of these services is available to compare users' detections of ovaries and follicles with the referential counterparts in the database. Sensitivity, precision, and accuracy of the compared detections were computed according to our protocol. We applied the same measures to verify inter-rater agreement, which indicates the consistency and reliability of the referential follicle segmentations as delineated by two independent raters. The USOVA3D database, with the established verification protocol, also allows us to compare the algorithms of

different research groups relevantly on the same data. Such a comparison in the field of Follicle Detection in Ovarian Ultrasound Volumes has not been possible so far (see [6], [7] and references within). Besides, our baseline algorithms confirmed that this database can be a reliable source for developing new detection methods.

Let us conclude this paper with some future work directions. The USOVA3D database size is rather limited in its first version, and we plan to expand this database to at least 100 volumes, annotated by a minimum of two raters. We are also going to encourage other researchers to contribute their segmented volumes of ovaries for possible inclusion into our database. The main criteria for inclusion will be compatibility with USOVA3D, and evidence that experienced raters segmented the images, the data were collected in the research approved by Ethical Committees, and informed Consents were signed by the participants. Later on, we would like to implement a so-called 'portable verification system' that the users could download to their computers. Such system will allow test validation locally, signing the results digitally, and uploading them to our portal.

Acknowledgement

This study was supported by the Slovenian Research Agency (Contract P2-0041).

References

[1] P. S. Hiremath, J. R. Tegnoor, Advancements and Breakthroughs in Ultrasound Imaging, IntechOpen Limited, 2013, Ch. Follicle Detection and Ovarian Classification in Digital Ultrasound Images of Ovaries, pp. 167–199. doi:dx.doi.org/10.5772/56518.

[2] V. Vlaisavljević, M. Došen, Clinical applications of ultrasound in assessment of follicle development and growth, Donald School J. Ultra. Obst. Gynec. 2 (1) (2007) 50–63.

[3] B. Ata, A. Seyhan, S. L. Reinblatt, E. Shalom-Paz, S. Krishnamurthy, S. L. Tan, Comparison of automated and manual follicle monitoring in an unrestricted population of 100 women undergoing controlled ovarian stimulation for IVF, Hum. Reprod. 26 (1) (2010) 127–133. doi:10.1093/humrep/deq320.

[4] T. D. Deutch, A. Z. Abuhamad, Sonography-based automated volume count (SonoAVC): An efficient and reproducible method of follicular assessment, GE Healthcare, Waukesha, USA (2007).

[5] T. D. Deutch, I. Joergner, D. O. Matson, S. Oehninger, S. Bocca, D. Hoenigmann, A. Abuhamad, Automated assessment of ovarian follicles using a novel three-dimensional ultrasound software, Fertil. Steril. 92 (5) (2009) 1562–1568.

[6] B. Cigale, D. Zazula, Directional 3D wavelet transform based on gaussian mixtures for the analysis of 3D ultrasound ovarian volumes, IEEE Trans. Pattern. Analy. Mach. Intel. 41 (1) (2019) 64–77. doi:10.1109/TPAMI.2017.2780248.

[7] B. Potočnik, B. Cigale, D. Zazula, Computerized detection and recognition of follicles in ovarian ultrasound images: a review, Med. Biol. Eng. Comput. 50 (12) (2012) 1201–1212. doi:10.1007/s11517-012-0956-y.

[8] O. R. Isah, A. D. Usman, A. M. Tekanyi, A hybrid model of PSO algorithm and artificial neural network for automatic follicle classification, Int. J. Bioautomat. 21 (1) (2017) 43–58.

[9] A. D. Usman, O. R. Isah, A. M. S. Tekanyi, Application of artificial neural network and texture features for follicle detection, African J. Comp. & ICT 8 (4) (2015) 2–9.

[10] D. S. Wanderley, C. B. Carvalho, A. Domingues, C. Peixoto, D. Pignatelli, J. Beires, J. Silva, A. Campilho, End-to-end ovarian structures segmentation, in: R. Vera-Rodriguez, J. Fierrez, A. Morales (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, LNCS, Vol. 11401, Springer International Publishing, 2019, pp. 681–689. doi:10.1007/978-3-030-13469-3_79.

[11] S. A. C. Marques, Ovarian structures segmentation using a neural network approach, Master's thesis, Faculdade de engenharia da universidade do Porto, Porto (2019).

[12] B. Potočnik, D. Zazula, Improved prediction-based ovarian follicle detection from a sequence of ultrasound images, Comput. Methods Programs Biomed. 70 (2003) 199–213.

[13] T. Chen, W. Zhang, S. Good, K. Zhou, D. Comaniciu, Automatic ovarian follicle quantification from 3D ultrasound data using global/local context with database guided segmentation, in: Proceedings of the 12th IEEE Conference on Computer Vision, Kyoto, Japan, 2009, pp. 795–802.

[14] D. Cheng, D. K. Prasad, M. S. Brown, Illuminant estimation for color constancy: Why spatial domain methods work and the role of the color distribution, JOSA A 31 (5) (2014) 1049–1058.

[15] Z. Emeršič, V. Štruc, P. Peer, Ear recognition: More than a survey, Neurocomputing 255 (13) (2017) 26–39. doi:10.1016/j.neucom.2016.08.139.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: Proceedings of 2010 IEEE Comp. Soc. Conf. Comp. Vis. Pattern Recog., San Francisco, USA, 2010, pp. 94–101. doi:10.1109/CVPRW.2010.5543262.

[17] P. A. Yushkevich, Y. Gao, G. Gerig, ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, USA, 2016, pp. 3342–3345.

[18] W. J. Schroder, K. M. Martin, L. S. Avila, VTK User's Guide – VTK File Formats, chapter 14, Kitware Inc., New York, USA, 2000.

[19] T. Boronczyk, E. Naramore, J. Gerner, Y. L. Scouarnec, J. Stoltz, Beginning PHP 6, Apache, MySQL 6 Web Development, Wrox Press, Birmingham, UK, 2009.

[20] B. Williams, D. Damstra, H. Stern, Professional WordPress: Design and Development, 3rd Edition, John Wiley & Sons, Indianapolis, USA, 2015.

[21] S. Šinjur, J. Munda, Iskanje podobnih video posnetkov kot spletna storitev v oblaku, in: Proceedings of the 28th Intern. Elect. Comput. Sci. Conf. ERK 2014, Vol. B, Portorož, Slovenia, 2014, pp. 64–67.

[22] S. Šinjur, D. Zazula, B. Žalik, Fast convex layers algorithm for near-duplicate image detection, Inform. 23 (4)

(2012) 645–663.

[23] W. Dormann, Google authentication risks on iOS, in: Mobile! 2016: Proceedings of the 1st International Workshop on Mobile Development, ACM, New York, USA, 2016, pp. 3–5. doi:10.1145/3001854.3001862.

[24] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union L 119/1 (2016) 1–88.

[25] B. Potočnik, D. Zazula, Assessing the efficiency of segmentation algorithms for ultrasound images, Electrotech. Rev. 68 (2-3) (2001) 97–104.

[26] T. Wiatowski, H. Bolcskei, A mathematical theory of deep convolutional neural networks for feature extraction, IEEE Trans. Inf. Theo. 64 (3) (2018) 1845–1866.

[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proceedings of Inter. Conf. Med. Img. Comput. Comp.-Assist. Interv. MICCAI 2015, Vol. III, Munich, Germany, 2015, pp. 234–241.

[28] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, Massachusetts, USA, 2017.

[29] M. A. Gore, P. L. Nayudu, V. Valisavljević, Prediction of ovarian cycle outcome by follicular characteristics, stage 1, Hum. Reprod. 10 (9) (1995) 2313–2319.

Table 1: Results published on efficient follicle detection methods and related datasets on real ovarian ultrasound patient data. 'N/A' denotes Not Available, 'px' pixels, and 'vx' voxels. Other denotations are explained in the accompanying text.

Reference	Year	Approach	2D/3D	Performance		Testing dataset			
				Measure	Value	# images	Image size (px/vx)	# raters	Annotated objects
Isah et al. [8]	2017	Active contours + PSO + NN	2D	Sensitivity Specificity	100% 96.8%	25	N/A	N/A	Follicles
Marques [11]	2019	CNN	2D	Sensitivity Specificity DSC	80% 73% 0.7 ± 0.2	107	192 × 200 to 192 × 620	1	Ovary, follicles
SonoAVC [5]	2008	Proprietary	3D	Sensitivity FMR v Match	52% 9% 90%	31	N/A	1	Follicles
Chen et al. [13]	2009	Probabilistic	3D	Sensitivity V Match S MAD	78% 80% 1.9 ± 1.4 mm	101	243 × 177 × 121 (median)	1	Ovary, follicles
Cigale et al. [6]	2019	Directional 3D DWT	3D	Sensitivity Specificity V Match S MAD	88% 85% 98% 0.3 mm	30	165 × 141 × 181 (mean)	2	Ovary, follicles

Table 2: Details about USOVA3D database and its entries (volumes).

# volumes	35
Age of women	33.4 ± 5.6 years (min 22, max 43)
Acquisition devices	Voluson 630 and 730, Medison Accuvix XQ
Volume size	$[101 \dots 229] \times [89 \dots 193] \times [115 \dots 247]$ voxels
Min volume	$149 \times 103 \times 115$ voxels
Max volume	$185 \times 193 \times 223$ voxels
Voxel dimension	0.2 mm × 0.2 mm × 0.2 mm
Training set	16 entries (predefined database splitting)
Entry	Volume + 2 ovary annotations (rater 1 and 2)
	+ 2 follicles annotations (rater 1 and 2)
Testing set	19 volumes (predefined database splitting)
Web link	https://usova3D.um.si

Table 3: Inter-rater variability of follicle annotations. Metrics for the selected rater were calculated by taking the follicle annotations of the unconsidered rater as the ground truth. Denotations are explained in Eq. (5).

	S	P	$\overline{\rho_1\rho_2}$	r_V	$\overline{e}$	$\overline{d}$	median(ξ_{vol})	min(ξ_{vol})	max(ξ_{vol})	ξ_{rater}
Rater 1	0.88	0.85	0.76	0.99	1.69	1.57	84.1	72.7	91.2	83.9
± std	0.15	0.25	0.08	0.01	0.48	0.65				
Rater 2	0.85	0.88	0.76	0.99	1.62	1.57	84.5	70.4	91.5	83.1
± std	0.25	0.15	0.08	0.02	0.44	0.65				

Table 4: Inter-rater variability of ovary annotations. Metrics were calculated by taking the ovary annotations of the unconsidered rater as the ground truth. Denotations are explained in Eq. (5).

	S	P	$\overline{\rho_1\rho_2}$	r_V	$\overline{e}$	$\overline{d}$	median(ξ_{vol})	min(ξ_{vol})	max(ξ_{vol})	ξ_{rater}
Rater 1	1	1	0.79	1	5.12	4.37	79.1	52.7	96.0	76.1
± std			0.12		3.16	3.46				
Rater 2	1	1	0.79	1	5.60	4.37	78.8	45.5	96.1	75.5
± std			0.12		4.02	3.46				

Table 5: Effectiveness of the proposed 'baseline' follicle detection methods evaluated on the USOVA3D database. Final score statistics and the overall algorithm score are presented.

Method	median(ξ_{vol})	min(ξ_{vol})	max(ξ_{vol})	ξ_{alg}
CNN	75.1	43.8	91.5	72.5
3D DWT	79.3	59.7	90.6	78.2

Table 6: Effectiveness of the proposed 'baseline' methods when detecting ovaries in ultrasound volumes. Final score statistics and the overall algorithm score are presented.

Method	median(ξ_{vol})	min(ξ_{vol})	max(ξ_{vol})	ξ_{alg}
CNN	73.6	40.5	87.9	72.2
3D DWT	72.5	18.3	87.1	63.3

Declaration of Competing Interest

The authors declare that there is no conflict of interest.