Guided Random Forests for Identification of Key Fetal Anatomy and Image Categorization in Ultrasound Scans

Mohammad Yaqub[1], Brenda Kelly[2], A.T. Papageorghiou[2], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, Engineering Science, University of Oxford
[2] Nuffield Department of Obstetrics & Gynaecology University of Oxford

Abstract. In this paper, we propose a novel machine learning based method to categorize unlabeled fetal ultrasound images. The proposed method guides the learning of a Random Forests classifier to extract features from regions inside the images where meaningful structures exist. The new method utilizes a translation and orientation invariant feature which captures the appearance of a region at multiple spatial resolutions. Evaluated on a large real world clinical dataset (~30K images from a hospital database), our method showed very promising categorization accuracy (accuracy$_{\text{top1}}$ is 75% while accuracy$_{\text{top2}}$ is 91%).

Keywords: Random Forests, Ultrasound, Image Categorization, Classification, Normalized Cross Correlation.

1 Introduction

Image categorization is a well-known open-research problem in computer vision for which many solutions have been proposed [1-3]. In medical image applications, image categorization is relatively under-investigated but nevertheless important. The volume of digital images acquired in the healthcare sector for screening, diagnosis or therapy is very large and increasing steadily.

Ultrasound based fetal anomaly screening is usually performed at 18 to 22 weeks of gestation. Several images of fetal structures are acquired following a standardized protocol. This screening scan aims to determine whether the fetus is developing normally by assessing several ultrasound images of different fetal structures. The number of different structures that are imaged and acquired in a complete scan varies, for example, the UK NHS Fetal Anomaly Screening Programme (FASP) recommends 21 views to be assessed and at least 9 images to be stored. The number of individual women undergoing a scan is often of the order of several thousands per department per annum. Most clinical departments save these scans to an archive system without any labeling of these images. This means it is not possible to recall images of body parts conveniently for later review or measurement nor to compare scans of the same fetus over time, or conduct automatic measurement post-acquisition.

Manual categorization is of course theoretically possible. However, it is expensive as it requires a good level of expertise alongside being tedious and time consuming. In this paper, we propose a method to automatically categorize fetal ultrasound images from anomaly scans. The new method is built on a machine learning classifier (Random

N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 687–694, 2015.
DOI: 10.1007/978-3-319-24574-4_82

Forests RF) in which we propose novel ideas to guide the classifier to focus on regions of interest (ROI). Although there are a number of methods which have been proposed to address different medical image categorization problems [4-6], very little research has been done in fetal ultrasound image categorization [4, 7, 8]. This may in part be explained because fetal ultrasound image categorization has unique challenges - the quality and appearance of the images vary for a number of reasons including variation of fetal position, sonographer experience, and maternal factors all affect the appearance of images. In addition, one fetal ultrasound image can contain one or more fetal and non-fetal structures. This implies that general classification methods applied to a whole image to output single-class fail to achieve this task robustly. Here we instead propose to restrict classification to candidate regions of interest.

The most closely related work to ours is [4] in which the authors proposed a machine learning method based on Probabilistic Boosting Tree (PBT) to automatically detect and measure anatomical structures in fetal ultrasound images. The method learns a PBT from coarse to fine resolution using Haar features to output object location. In our work, we also use a family of Haar features but ours are generalizable and allow capturing not only local but global appearance. In addition, the ultimate goal of [4] was to perform measurements on the detected object. Our goal is the more general goal of categorizing images by their content. Moreover, we evaluate our method on a much larger dataset (30K vs 5K). Finally, we base our solution on a RF classifier which has proven to outperform many state-of-the-art learners including boosting methods [9].

2 Guided Random Forests

In general, the proposed method first extracts novel features from 2D fetal ultrasound images to build a RF classifier which learns the anatomy within any image. The classifier learns eight classes which represent seven anatomical views 1) head in the trans-ventricular plane (Head TV), 2) head in the trans-cerebellar plane (Head TC), 3) 4-chamber view of the heart (cardio), 4) face (a coronal view of the lips), 5) abdomen (at the level where the abdominal circumference is measured), 6) femur (for the measurement of femur length), 7) spine and 8) a category called "Others" which may contain many other fetal structures e.g., limbs, kidneys, cord insertion, etc.

Learning is guided because the proposed features are extracted from regions where structures of interest exist. This helps avoid misleading anatomy within these images. We propose a method to build the features computed on these regions in a translation, orientation, and scaling invariant fashion that is key to make our proposed learning algorithm robust.

2.1 Localizing Object of Interests

The main objective in this method is to automatically localize a single object of interest. The proposed RF method samples features from within discriminative regions instead of looking blindly everywhere across the images. As noted earlier, this is important in ultrasound where there can be distracting regions of similar appearance. To achieve this we have built several multi-resolution and multi-orientation geometric templates which capture the geometric appearance of the main structures we are interested in classifying.

To find the best match between an image I and a template T, we compute the normalized cross-correlation between I and T in the Fourier domain. The parameters for the image region which produces the maximum correlation with the template are recorded. These 9 parameters are $\{corr, s, d, \underline{C}, \underline{P_1}, \underline{P_2}\}$ such that $corr$ is the template correlation value (response), s is the relative size of the matched region with respect to the original image size, d is the Euclidean distance measured between the center of the region $\underline{C} = \{cx, cy\}$ and the center of the image, $\underline{P_1} = \{x_1, y_1\}$ is the location of the top left corner of the matched region within the image, and $\underline{P_2} = \{x_2, y_2\}$ is the location of the bottom right corner.

We utilize established clinical knowledge of fetal anatomy in the templates to ensure that their physical size correspond to the size of fetal structures. For instance, biparietal diameter (the minor axis of the fetal skull) of the fetal skull ranges from 40mm to 55mm [10]. Fig. 1 shows an example from each template type. Fig. 2 illustrates the matched regions of best-matched templates on three ultrasound images (see supplementary videos for further examples). We use *green* for skull, *blue* for spine, *red* for face, *magenta* for abdomen and *yellow* for femur. The width of the border of the box indicates the strength of the template response (*corr*).

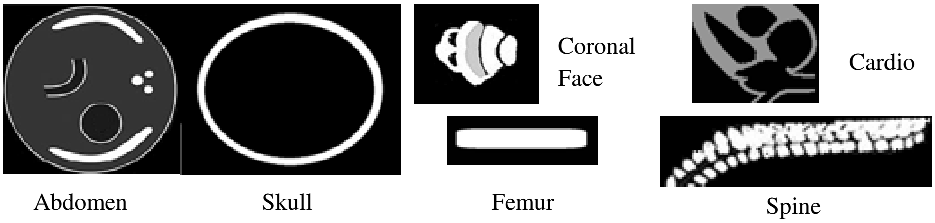
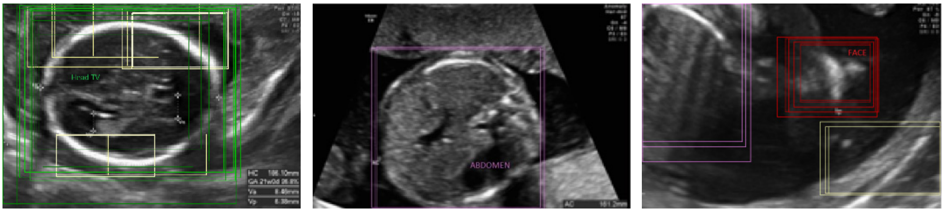


Fig. 1. Examples of a representative set of fetal structure templates.



(a) Head TV (b) Abdomen (c) Face

Fig. 2. Template responses on three images (regions with high responses are only shown here for convenience). The template response is depicted as a colored rectangle (see text for color definition).

2.2 Feature Sets for the Learning Algorithm

Our machine learning solution utilizes two feature sets. The first feature set represents template metadata ($corr$, s, d) as defined in Section 2.1. These values are highly discriminative since $corr$ provides an indication of how probable it is that a region contains a structure of interest, s indicates how large a region is within an image e.g., fetal face occupies a small region within an image while a skull is typically larger, and d is the Euclidian distance between a region and the center of the image i.e., in sonography the structure of interest typically appears in the middle of the image; therefore, if an image contains two anatomical structures, this feature will lean to classify the image according to the structure that appears closer to center of the image. The number of features in this feature set is small (77×3=231; here 77 is the number of templates) but highly discriminative between some structures e.g., head TV vs spine. However, this set is not rich enough alone for good discrimination. For instance, the skull template will not distinguish between head TV and head TC and the femur and spine templates look similar and can provide similar responses.

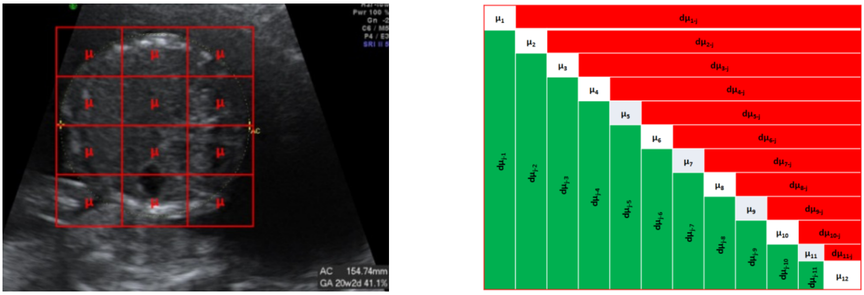
The second feature set captures the appearance of a region R by splitting the region into a number of rows r and columns c such that r and c are randomly selected by the learning algorithm. This random splitting allows multiresolution distinction between regions (scaling invariance). In other words, if r and c are small, global appearance of the region is captured while if r and c are as large as the number of rows and columns in the region then fine detail appearance is learned. Fig. 3 (a) shows a candidate region which is split into 12 blocks. In general, the mean intensity is computed inside each block $k : [1, (r{\times}c)]$. To compute the relative appearance between blocks within a region, a square feature matrix A (as in Fig. 3 (b)) of $(r{\times}c)$ rows and columns is computed as follows

$$\forall \begin{bmatrix} i = 1:(r \times c) \\ j = 1:(r \times c) \end{bmatrix}, \quad A_{k_i,k_j} = \begin{Bmatrix} Mean\left(block_{k_i}\right) & \mid i = j \\ Mean\left(block_{k_i}\right) - Mean\left(block_{k_j}\right) & \mid i \neq j \end{Bmatrix}, \qquad (1)$$

As in [11], a feature value is then computed as

$$f = log_e\,(det(A)) \quad : \quad det \text{ is the determinant of a matrix} \qquad (2)$$

Because the location of a region is not fixed within the image, the computed feature over a region is translation invariant. In addition, because the matrix A represents the variance between a given block and the rest irrespective of block position within a region, this feature type robustly learns region appearance at different orientations.



(a) splitting the region into random number of blocks

(b) Feature matrix A (12×12 in this example). *Green area* is mean(block$_i$) – mean(block$_j$) while *red* is the negative of it.

Fig. 3. An example region of 4×3 blocks on a fetal abdominal image and the feature matrix A.

2.3 Training and Testing

Before training and testing, all images are resized to the same pixel spacing to simplify correspondence between appearance features. We have developed a classifier based on RF which learns a fetal image category from a set of fetal ultrasound anomaly images. Training is performed on a set of images $\mathcal{I}=\{I_1, ..I_N\}$ such that N is the total number of images and their class labels $\mathcal{L}=\{L_1, ..L_N\}$ such that the ith training example is parameterized by $\{I_i, \{f_1, f_2, \ldots f_F\}, L_i\}$ where f is a feature from the feature pool of size F. The classifier learns the best combination of the features described in Section 2.2 to build a set of trees. Each tree node is created out of the best feature/threshold from a set of randomly sampled features from the feature pool. Once the best feature and threshold are found for a tree node, the list of training examples branch left and right and the training proceeds recursively on each subtree. A leaf node is created if the maximum tree depth is reached or if the number of training examples of a node is small. The set of training images reaching a leaf is used to create a distribution which can be used to classify unseen images during testing.

3 Experiments

3.1 Dataset

29858 2D fetal ultrasound images from 2256 routine anomaly scans undertaken in the 12 months of 2013 between 18 weeks + 0 days to 22 weeks + 6 days were used from a hospital database. These scans were acquired by 22 different sonographers so there is large (but typical) variation in image quality, machine settings, and zoom aspects. All images were manually categorized by 14 qualified sonographers. No attempt was made to reject or remove any scans from the dataset as we intended to evaluate our method on a real-world dataset. This means the reported results show potential performance in real world deployment. The dataset contained fetal head (TV and TC), abdomen, femur, spine, face, cardio, and several other categories as specified in the standard screening protocol in the hospital. Scans were acquired on a GE Voluson E8 machines (GE Healthcare, Milwaukee USA). All images were anonymized before analysis, and their use underwent institutional board approval.

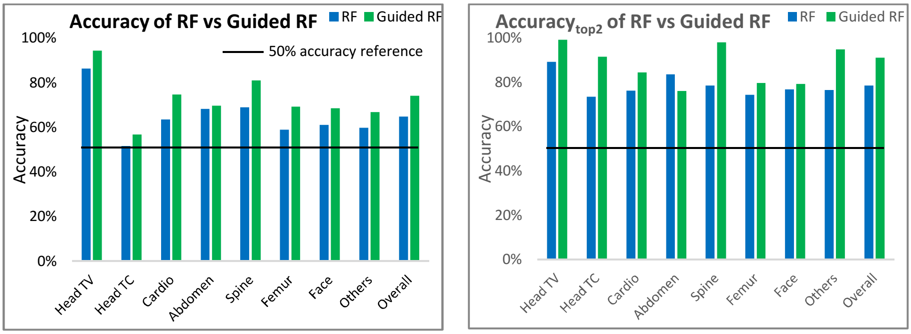
3.2 Evaluation Protocol and Metrics

The accuracy of the template matching was visually assessed, see supplementary material. We investigated the effect of the new proposed features by comparing the guided RF with a traditional RF using standard generalized Haar wavelets used in many medical image applications including [11-16]. These features compute the difference of mean intensity of two random blocks. The whole image was considered when sampling features and the center of the image used as a reference point for those features. In addition, we investigated the categories that found confusing to gain further understanding of the algorithm performance.

We have performed 10-fold cross validation and we have selected RF parameters experimentally. These were then fixed in all reported experiments. Maximum tree depths of 18 and 30 trees were trained. RF produces probabilities during testing. An image is correctly classified (defined by the true positives *TP*), if its class has the maximum probability among all other class probabilities. We report the accuracy of a method as ($acc = TP\ /\ N$). However, due to the complexity of fetal images and the occurrence of multiple structures in many images, we also report $\mathrm{acc}_{\mathrm{top2}} = TP_{top2}\ /N$ (as used in many machine learning imaging papers e.g., [17, 18]). An image is considered in TP_{top2} if its class is within the top 2 probabilities of the algorithm. Finally, all experiments were implemented in Matlab and all RF methods were trained and tested in a parallel manner to achieve fast implementation. Given an unseen image, Guided RF categorizes it in approximately 0.32sec on a high end workstation with 20 processor units.

4 Results

Fig. 2 shows visual results of localization of objects of interest where boxes are drawn around the matched regions for the different templates. Only the best-matched templates are shown. Each color represents a different template type as described in Section 2.1 and the width of the box border signifies the strength of the template response. Further typical results can be found in the supplementary material. The overall accuracy of the traditional RF method was 65% while Guided RF achieved 75% accuracy. A detailed comparison on the different classes is shown in Fig. 4 (a). Fig. 4 (b) shows the accuracy$_{\mathrm{top2}}$ result for both the traditional RF and the Guided RF. Guided RF increased to 91% accuracy when considering the top 2 probabilities from RF output while traditional RF increased to 79% accuracy. Finally, Fig. 5 shows the category confusion plot which presents the different categories in descending order with respect to how often images get classified in. Note that the most common error is for one of the eight classes to be mis-classified as “Others” as opposed to being confused with another named fetal class.



(a) Accuracy of both RFs on the top probability (b) Accuracy of both RFs on the top 2 probabilities

Fig. 4. Accuracy of the proposed method on the different image categories.

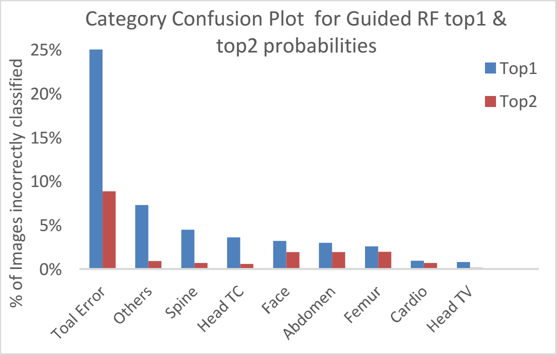


Fig. 5. Category Confusion Plot. From the 8 categories, this figure shows the misleading categories in descending order for the top1 categorization. Confusing categorizes is also shown for Top2. For instance, when a "Head TV" gets categorized as "Others" then the "Others" gets one more confusing case.

5 Discussion and Future Work

We have presented a new machine learning solution based on RF to categorize fetal ultrasound images. We showed how the RF classifier can be guided to provide good classification via guided sampling of features from ROIs. We first presented a method to detect ROIs using template matching to allow the classifier to learn from these regions and ignore misleading regions. We also presented a new feature type which captures a region within an image such that this feature is translation and orientation invariant. This type of feature proved to be important in our application because the position and orientation of fetal structures depends on the location of the fetus in the womb which is highly variable, see Fig. 4.

Because of the complexity of the "Others" group and the existence of many images in this group which look similar to the other 7 groups (e.g., cord insertion images are very similar to abdominal images), many images gets sorted as "Others". This can be seen in Fig. 5. Also, note that the "Spine" group may also be confusing as its appearance is similar to different structures, e.g., femur and diaphragm which is in the "Others" group. A multiclass solution which accommodates context of other labels may solve this problem and will be investigated in future work. Finally, while we have considered 2D ultrasound, the method is generalizable and could be extended to 3D ultrasound and other imaging modalities.

References

[1] Murray, R.F.: Classification images: A review. Journal of Vision 11 (2011)

[2] Bosch, A., Muñoz, X., Martí, R.: Review: Which is the best way to organize/classify images by content? Image Vision Computing 25, 778–791 (2007)

[3] Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28, 823–870 (2007)

[4] Carneiro, G., Georgescu, B., Good, S., Comaniciu, D.: Detection and Measurement of Fetal Anatomies from Ultrasound Images using a Constrained Probabilistic Boosting Tree. IEEE Transactions on Medical Imaging 27, 1342–1355 (2008)

[5] Prasad, B.G., Krishna, A.N.: Classification of Medical Images Using Data Mining Techniques. In: Das, V.V., Stephen, J. (eds.) CNC 2012. LNICST, vol. 108, pp. 54–59. Springer, Heidelberg (2012)

[6] Greenspan, H., Pinhas, A.T.: Medical Image Categorization and Retrieval for PACS Using the GMM-KL Framework. IEEE Transactions on Information Technology in Biomedicine 11, 190–202 (2007)

[7] Maraci, M., Napolitano, R., Papageorghiou, A., Noble, J.A.: Object Classification in an Ultrasound Video Using LP-SIFT Features. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Müller, H., Zhang, S., et al. (eds.) Medical Computer Vision: Algorithms for Big Data (2014)

[8] Ni, D., Yang, X., Chen, X., Chin, C.-T., Chen, S., Heng, P.A., et al.: Standard Plane Localization in Ultrasound by Radial Component Model and Selective Search. Ultrasound in Medicine & Biology 40, 2728–2742 (2014)

[9] Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 15, 3133–3181 (2014)

[10] Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., et al.: International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. The Lancet 384, 869–879

[11] Criminisi, A., Shotton, J., Robinson, D., Konukoglu, E.: Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In: Presented at the MICCAI Workshop in Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging, Beijing, China (2010)

[12] Chykeyuk, K., Yaqub, M., Noble, J.A.: Novel Context Rich LoCo and GloCo Features with Local and Global Shape Constraints for Segmentation of 3D Echocardiograms with Random Forests. In: Menze, B.H., Langs, G., Lu, L., Montillo, A., Tu, Z., Criminisi, A. (eds.) MCV 2012. LNCS, vol. 7766, pp. 59–69. Springer, Heidelberg (2013)

[13] Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A.: Neighbourhood Approximation Forests. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 75–82. Springer, Heidelberg (2012)

[14] Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Möller, A.M., Nekolla, S., Navab, N.: Fast Multiple Organs Detection and Localization in Whole-Body MR Dixon Sequences. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 239–247. Springer, Heidelberg (2011)

[15] Yaqub, M., Javaid, M.K., Cooper, C., Noble, J.A.: Investigation of the Role of Feature Selection and Weighted Voting in Random Forests for 3-D Volumetric Segmentation. IEEE Transactions on Medical Imaging 33, 258–271 (2014)

[16] Yaqub, M., Kopuri, A., Rueda, S., Sullivan, P.B., McCormick, K., Noble, J.A.: A Constrained Regression Forests Solution to 3D Fetal Ultrasound Plane Localization for Longitudinal Analysis of Brain Growth and Maturation. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 109–116. Springer, Heidelberg (2014)

[17] Alex, K., Sutskever, I., Geoffrey, E.H.: ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 1097–1105 (2012)

[18] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al.: ImageNet Large Scale Visual Recognition Challenge. Computer Vision and Pattern Recognition, arXiv:1409.0575 (2014)