Comparison of Texture Synthesis Methods for Content Generation in Ultrasound Simulation for Training

Oliver Mattausch[a], Elizabeth Ren[a], Michael Bajka[b], Kenneth Vanhoey[a], Orcun Goksel[a]

[a]Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland;
[b]Departments of Obstetrics and Gynecology, University Hospital Zurich, Zurich, Switzerland

ABSTRACT

Navigation and interpretation of ultrasound (US) images require substantial expertise, the training of which can be aided by virtual-reality simulators.[1] However, a major challenge in creating plausible simulated US images is the generation of realistic ultrasound speckle. Since typical ultrasound speckle exhibits many properties of Markov Random Fields, it is conceivable to use texture synthesis for generating plausible US appearance. In this work, we investigate popular classes of texture synthesis methods for generating realistic US content. In a user study, we evaluate their performance for reproducing homogeneous tissue regions in B-mode US images from small image samples of similar tissue and report the best-performing synthesis methods. We further show that regression trees can be used on speckle texture features to learn a predictor for US realism.

Keywords: Ultrasound simulation, speckle, texture synthesis

1. INTRODUCTION

Ultrasound (US) is a low-cost, real-time, and radiation-free medical imaging modality. However, due to its low signal-to-noise ratio and characteristic artifacts, its interpretation requires substantial expertise. Additionally, since it generates images in real-time but often merely a plane into 3D anatomy, US transducer navigation is a challenging topic. As reported by Reis et al.,[2] aspiring sonographers are often unprepared for relatively common pathologies like heart failure, as they appear only in a fraction of cases and may have never be encountered by a medical student during a limited number of training sessions. To aid in training of medical personnel for these tasks, virtual-reality simulation is of significant interest. This will enable the generation of arbitrary patient anatomy and even rare pathologies, while also providing objective evaluation of training. To be of value, a US training simulation needs to present images with sufficient realism. A major challenge for realism is the generation of ultrasound speckle, which is the noise-like interference pattern of echos from countless sub-wavelength scattering sources in tissue (e.g. muscle fibers, cell nuclei, large proteins, etc), see Figures 1 and 3. Such scattering sources in artificial media (phantoms) can be parametrized well using algebraic models, e.g., a Gaussian distribution. However, US speckle in real tissue is not necessarily random (due to subtle local changes in tissue structure) but it often exhibits certain structure and visual richness that algebraic approximations cannot easily capture. Hence, tissue speckle could only be reproduced with limited realism in previous model-based simulations.[3]

Due to the convolutional nature of US image formation, US speckle exhibits many properties of Markov Random Fields (MRF). In computer graphics, exemplar-based texture synthesis methods,[4] many of which rely on a MRF assumption, allow to generate texture content that resembles an exemplar. We therefore hypothesize that *texture synthesis* can be adopted for generating plausible virtual US speckle from samples of existing B-mode images. The goal of this paper is to investigate different classes of texture synthesis methods to identify those that best simulate US speckle. Note that there is no established objective metric on speckle appearance, so we conducted a user survey with US imaging experts comparing different texture synthesis techniques. We hope our results to facilitate various simulation tasks, such as the inpainting of arbitrary anatomical models with realistic US speckle using previous samples of tissue appearance. During development, we do not have constant access to medical experts in order to select the best possible texture methods or their best parametrizations. Also, during content generation, a medical expert cannot always be present either. Having a user study organized each time

Corresponding author: Orcun Goksel (ogoksel@vision.ee.ethz.ch)

to access new method or parameters is a tedious approach. Therefore, we envision in the long run a way to "replace" medical experts in speckle appearance assessment. In an attempt to establish objective metrics, we additionally evaluated quantitative texture metrics on the images seeking for a linear or machine-learning based regression with subjective user evaluations.

Our contributions include (*i*) the design and execution of a perceptual study on subjective US speckle quality, (*ii*) the comparison of several texture synthesis methods for US speckle generation, and (*iii*) a study on potential objective metrics for predicting US texture quality.

2. RELATED WORK

In this work, we are interested in *simulation of US imagery*; *texture synthesis* methods, which we exploit to synthesize the former; and *texture similarity* metrics, which we exploit for predicting synthesis quality. We briefly review these below.

2.1 Ultrasound Image Simulation

A well-established model for generating ultrasound speckle is convolution-based ultrasound simulation.[5–7] In this model, based on the linear Born approximation of the full wave model, ultrasonic speckle intensity is obtained by convolving point-like *scatterers* in the tissue with the ultrasonic impulse response, also called the point-spread function (PSF). This model can produce realistic results, when the input scatterers are sufficiently realistically distributed. However, a realistic scatterer realization is not straight-forward to set. Earlier work proposed parametrizations of scatterer distributions from given tissue examples,[8,9] but such simplistic parametrizations are often insufficient to capture the variation in real ultrasound. Alternatively in this paper, of interest are thus non-parametric, learning, or example-based methods, of which several dedicated methods exist.[10] To our knowledge, there exists no survey reviewing the use of the wide array of methods available from other fields like computer graphics or vision for the purpose of ultrasound simulation.

2.2 By-example texture synthesis

By-example texture synthesis is the process of generating an arbitrarily large texture (called *output*) from an input image (called *exemplar*).[4] Most methods are based on a MRF assumption, i.e., synthesis is the realization of a *local* and *stationary* process. Locality means that synthesized pixel intensities/colors depend on their locally neighboring ones only. Stationarity refers to this being true for all pixels: There is no global variation in the image, i.e., no two pixels with similar surrounding can have a different intensity/color. A MRF-based synthesis method thus ensures local consistency: Ideally, any local neighborhood found in the output image should be present somewhere in the exemplar image. Local consistency can be expressed directly on pixel colors, or rather on pixel statistics (e.g., higher order correlations). The methods that implement this principle can be categorized into parametric ones (statistics-based) and non-parametric ones (pixel-color based). Among them, we chose four representative methods in this study.

Parametric texture synthesis methods define a model by a set of statistical parameters over the texture. Parameters are learned from the exemplar, and a sample is drawn from the model to create a new texture. This is implemented by starting from a white noise image, then iteratively adapting the synthesized texture until it matches the prescribed set of statistics. That set can be composed of pixel histograms and wavelet statistics,[11] or oriented wavelet filter responses at different scales.[12] Variety is ensured by a random white noise image initialization. However parametric methods are less used because of their theoretical complexity and unpredictability. Only recently, deep networks extended the state of the art in parametric texture synthesis,[13] where the model consists of learned statistics on the parameters of a convolutional neural network. These are out of the scope of this work. We studied the state of the art non-deep parametric method (PARAM).[12]

Non-parametric texture synthesis[14] starts from an empty output texture, then sequentially copies pixels from the exemplar to the output, up to completion of the latter. In that process, local coherence is ensured: The color of a new pixel is such that it is consistent with its neighbors, i.e., a similar neighborhood (of a predetermined size) exists in the exemplar. A drawback of this method is its low speed due to its sequential *one-pixel-at-a-time* synthesis process. Image analogies[15] extend it to multiple sizes of neighborhoods so that features are matched at multiple scales. A fast variant of the algorithm of Efros et al.[14] is Natural Texture Synthesis,[16] which improves coherence by preferring verbatim copies of an input region. Quilting[17] sequentially copies whole neighborhoods at a time and makes them slightly overlap in the output. The core search resides in finding the minimal overlap error. Additionally, a greedy algorithm finds the minimal cut on the overlapping areas, i.e., the patches are cut out where the visual error is minimal. While the above are sequential greedy algorithms, Texture Optimization[18] minimizes a global energy term composed of the sum of mismatches between output pixels and their best-matching neighborhoods in the exemplar. An expectation-maximization (EM) algorithm optimizes this energy by alternately improving the matching and the output pixel values.

We studied three non-parametric texture synthesis methods, each representing an important class of algorithms:[4] greedy pixel-copy (PIXEL)[14] (with a 5-pixel neighborhood), greedy patch copy (PATCH)[17] (with a 16 pixel patch size and 8 pixels of overlap), and global optimization[18] (OPTIM).

Other methods like the multi-scale extension[15] of PIXEL as well as its fast variant, Natural Texture Synthesis,[16] were also considered in an initial stage of our study. As these methods produced similar but consistently worse results than the PIXEL algorithm when applied to US imagery, we decided to discard them. This resulted in precious time gain in our subjective study that has to be deployed among medical experts.

2.3 Texture similarity metrics

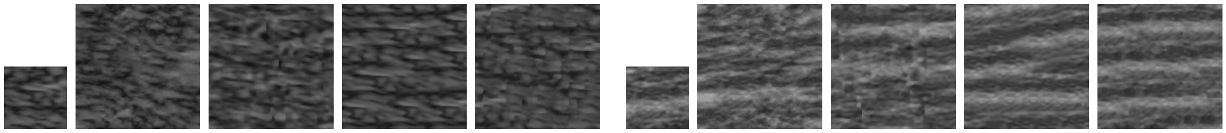
We aim to judge how similar a simulated/synthesized speckle image (e.g., Figure 1, large images) is to the actual ultrasound image (exemplar, e.g., Figure 1 small images) of the region it represents. Accordingly, we use several similarity metrics and texture feature distances. We investigate whether and which such similarity distances best encode and represent the realism of simulated speckle patches." Based on such feature distances, we later investigate the potential of automatic predictors, and train such a predictor to assess the realism of synthesized ultrasound speckle textures.

Image similarity metrics. Measuring similarity of textures is a hard task: two textures can be perceived similar while their pixel-to-pixel distance is high. Image similarity measures exist: they either simulate the low-level mechanisms of the human visual system[19] (HVS) or are based on intuitive hypotheses of how the HVS functions. These methods were validated earlier on natural images, not textures, let alone US imagery, so their applicability herein is uncertain. The structural similarity index (SSIM)[20] is known to be one of the methods that is most widespread and correlated to human perception. We include SSIM as a metric for our study.

Texture features. Texture features quantify visual concepts such as roughness, regularity, and visual similarity. In this work, we study several image and texture-based metrics known in vision and/or graphics for predicting the qualitative rating of experts in Section 3:

- *Textons* are a list of learned local texture filters used in recognition and classification tasks.[21] Each pixel is assigned its most representative texton. To compare exemplar and output, we use the distance between their respective histograms of texton assignments[22] in their rotationally-sensitive and -invariant versions.
- *Synthesizability* of a texture[23] predicts how well a particular image can be used as an exemplar for texture synthesis methods, i.e., how capable image synthesis techniques are in resynthesizing a particular texture. It is measured by a random forest trained with 7 texture features: 3 widely-used ones (Local binary patterns[24] (LBP), Schmid filter bank responses[25] (SFilter), GIST features[26]) and 4 ad-hoc ones specifically designed to measure "textureness", "homogeneity", "repetitiveness", and "irregularity". As our purpose is similar, we investigate them all. To obtain a distance between two images, we use the L_2-norm of their respective feature differences.
- *Grey-level co-occurrence matrices (GLCM)*[27] encode frequential information, i.e., the average frequency at which pairs of gray-levels occur for a list of spatial displacements. Features can be deduced from that information. We deduce "homogeneity", "energy", and "contrast" and subsequently apply a L_1-norm on exemplar and output feature differences.
- *Bhattacharyya* distance[28] is also used between the intensity histograms of the exemplar and output images.

Note that while textons and the synthesizability features have proved useful in computer graphics, the GLCM, LBP, and Bhattacharyya measures are more related to classification and recognition tasks in computer vision.



(a) Tissue-mimicking phantom (b) In-vivo muscle tissue

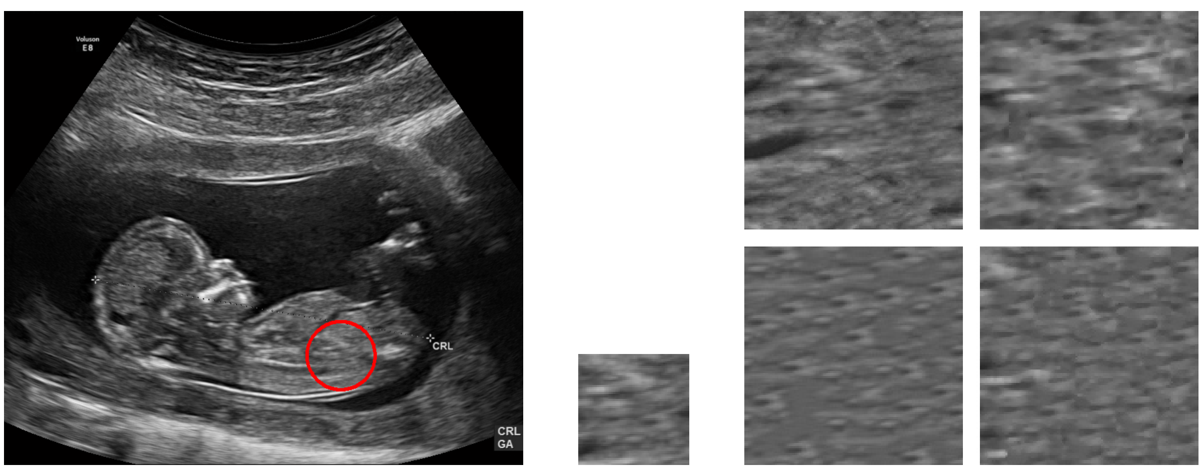
Figure 1. Texture synthesis applied to US imagery: a homogeneous US patch (smallest images) forms the exemplar from which a larger texture is synthesized. From left to right, the four techniques we considered: PIXEL, PARAM, OPTIM and PATCH.

3. METHODS

In this work, we wish to evaluate the visual quality of the example-based texture synthesis techniques PARAM, PIXEL, PATCH and OPTIM presented in section 2. These techniques all follow the MRF assumptions: locality and stationarity. To isolate the effect of texture synthesis methods, we have not considered additional techniques, e.g., those that take global variation into account, i.e., non-stationarity responsible for larger scale changes in tissue such as between two different anatomical structures. Such methods are often quite application-specific.[29] Accordingly, our study focuses on investigating the synthesis of homogeneous US image regions. These are also the areas where (convolutional) speckle appearance is isolated from large specific features from reflecting US echos at anatomical interfaces. In particular, given a small sample of an US B-mode region (the exemplar), the methods generate a larger region synthesized from it (the output), see Figure 1. In the following, we explain how we conducted our subjective user study.

3.1 Synthesized ultrasound image patches

We selected 11 US images of different anatomical structures, including liver, uterus, breast, muscle, and a fetus in the womb, as well as a tissue-mimicking US phantom of the uterus (404A, CIRS, Norfolk, VA). From these images, anatomically homogeneous regions were selected, from which a crop was made to form an exemplar image. Finally, these exemplars were fed to the 4 selected texture synthesis methods, producing $4 \times 11 = 44$ synthesized images. Our process is illustrated in figure 2.

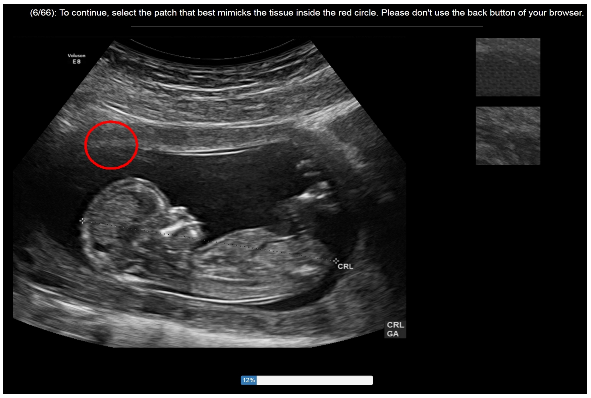


(a) Source US image and region (b) Crop (c) Four synthesized textures

Figure 2. From a US image, we identify a homogeneous region (red circle) and take a crop out of it. That crop is in turn used as an input exemplar for the different texture synthesis techniques. Each crop is used to synthesis four different images, one per synthesis technique.

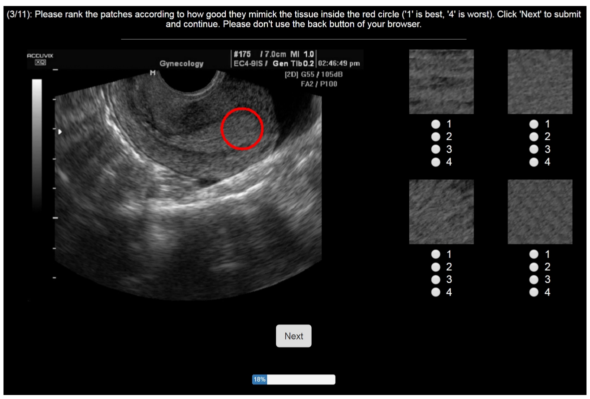
3.2 Experimental procedure

We would like to assess if synthesized US texture can well mimic the tissue it is supposed to represent, i.e., the anatomical tissue the exemplar was taken from. Therefore, we made several key study design choices that were

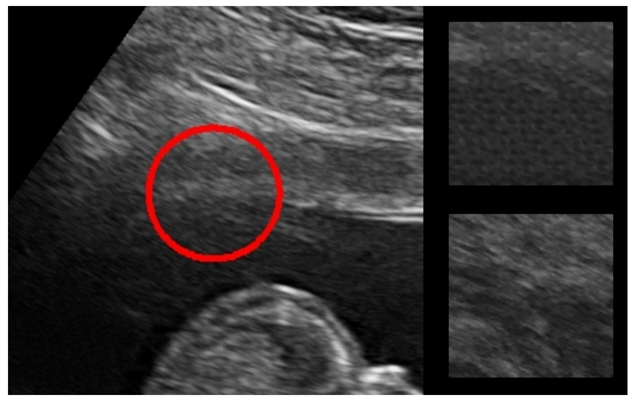


(a) Paired Comparison

(b) Rating



(c) Ranking



(d) Closeups of Paired Comparison

Figure 3. A question page of each of our three ultrasound questionnaires (a,b,c) and a zoom into the regions of interest and the generated texture patches for Paired Comparison (d).

found to be essential after conducting a preliminary study with 9 participants and several design iterations. First, we used a circular marker (red circle in the images) for the region for comparison, as opposed to a rectangular one, and we synthesize a larger region than the cropped exemplar. This was to prevent the participant from trying to do pixel-by-pixel matching between exemplar and output. Our goal is indeed not to generate the missing piece that best fits, but to generate a US image patch that best mimics the same underlying anatomical tissue. Second, we used a full-reference comparison method by providing the entire US image (not only the small homogeneous texture), since we wanted the expert user to be fully aware of the US context including the surroundings of the marked exemplar region. Finally, we only included US experts in our study for interpreting US imagery.

We wanted to collect subjective evaluations on all 4 methods for all the 11 exemplars. There are several different question and questionnaire styles that one can follow. Considering complementary information that can be gained from different question formats, we conducted three separate questionnaires with each user: paired comparison, ranking, and rating, respectively (see figure 3). These all judge speckle realism in a similar but slightly different way. All 3 tests use full-reference comparisons, i.e., images are compared to a reference image region (exemplar). The question types were:

Paired comparison. "Select the patch that best mimicks the tissue inside the red circle."

Users are asked to choose the most realistic patch with respect to a reference image (Figure 3(a)). This is the simplest mode of interaction, with a low cognitive workload, hence traditionally results in the most reliable rankings.[30] However, many pairs have to be shown to the user so as to obtain a complete comparison of all patch pairs: $(n-1)!$, where n is the number of stimuli to compare. For $n = 4$ for us, we showed all $3! = 6$ synthesized patch pairs for all 11 images, hence 66 comparisons in total for each user.

Rating. "On a scale from '–' (implausible) to '++' (plausible), please rate how well a patch mimicks the tissue inside the red circle."
The users were asked to rate each single synthesized patch on a 5-point Likert scale (from '–' to '++', see Figure 3(b)). This task is also simple, but it is not a comparison but rather a judgment of the implicit value (realism) of each patch according to the observer. In this scenario, n questions are asked per exemplar, thus $4n = 44$ questions in total for each user.

Ranking. "Please rank the patches according to how well they mimick the tissue inside the red circle."
The users are asked to rank all n stimuli w.r.t. each other (Figure 3(c)). This is similar to the comparison task, but where the user presented with all potential synthesis options at once. So, the user is forced to make a sorted preference for the methods. This ranking is somewhat more tedious for the user than the comparisons, but it requires 1 question page per exemplar, thus only 11 questions in total for each user.

Note that the comparisons and ranking are relative evaluations of different methods with respect to each other, whereas rating requires the user to make a subjective judgment on the realism of US appearance, i.e. even the best ranked method might turn out to have insufficient realism for simulation.

3.3 Technical details.

Exemplar textures are cropped within the red circle and are of size 64×64 pixels. Synthesized patches are of size 128×128, thus double the size of exemplars. Note that if the synthesized patches being larger also prevented the example-based methods choosing (synthesizing) the entire patch by only using the exemplar itself: This would create perfectly realistic images, but would not be suitable for any use-scenario where larger or different-shaped regions are to be filled.

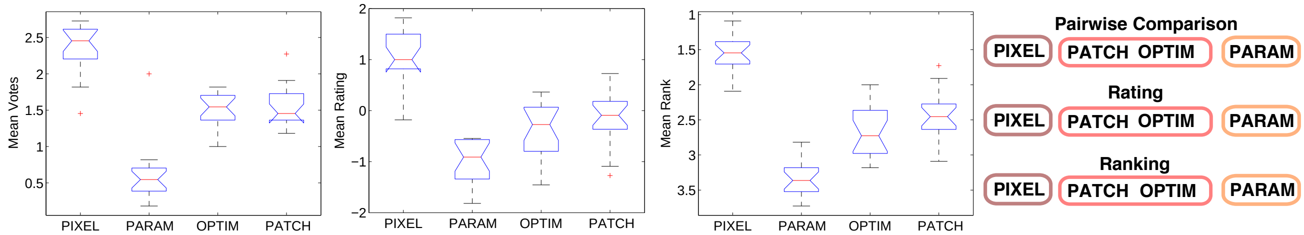
Dissemination of the study was implemented as an web-based tool that can run in a browser. This made distribution to experts (who often have limited time) convenient, and it is easy to extend in the future to larger groups or to larger image/method datasets. We disseminated participant-specific links, where the user responses were recorded automatically by the web-server at the end of each three questionnaires. All users were asked to complete the experiments in the order: paired comparison, rating, ranking. The order of the reference images within the questionnaires as well as the positions of the synthesized textures were randomized between participants, to avoid any bias to the order of appearance or placement. The first 3 questions of the comparison questions were given as training examples to the user at the beginning of Paired Comparison, the results of which were discarded. These questions were then asked again at the end of this questionnaire, where they were recorded, which then practically increased the number of comparison questions to 66+3=69.

Prior to the questionnaires, users were asked information on their background and expertise, namely: gender, medical specialization (one of "obstetrics and gynecology", "internal medicine", "surgeon", "radiologist", "anesthetist", "medical student", "technical expert" or "other"), expertise (in terms of number of performed US scans, among "less than 200","between 200 and 1000", or "more than 1000"), and fitness (in terms of number of US scans performed per week).

4. RESULTS

19 ultrasound experts participated in our study and finished the three questionnaires, 12 male and 7 female. 9 participants were specialized in Obstetrics/Gynecology, 1 was a radiologist, and 9 were technical experts. With respect to experience, 8 participants performed less than 200 ultrasound scans in their lifetime, 2 participants between 200 and 1000, and 9 performed over 1000 scans. The average number of ultrasound scans performed per participant was 18.1 per week, with a standard deviation of 24.7. The average time spent per questionnaire was 8.7 min for Paired Comparison, 8.3 min for Ranking and 6.0 min for Rating.

Synthesizing the textures happened in a pre-computation step, hence timings are not critical. Nevertheless, faster synthesis is more convenient. The timings required for synthesizing a single texture were 0.8 sec for PATCH, 10.5 sec for PIXEL, 115 sec for PARAM, and 326 sec for OPTIM on a Lenovo Thinkpad W540 with a Intel i7-4900 CPU (using a single core) and 24 GB memory.



(a) Paired Comparison (b) Rating (c) Ranking (d) Post-hoc analysis

Figure 4. (a)-(c): The results of the 3 questionnaires. Note that for Ranking, a lower rank is better. PIXEL performs significantly better than the other methods, as shown by a statistical significance analysis (Tukey's HSD) (d).

4.1 Subjective quality estimation

Figure 4(a-c) show the means and variances of scores for the different subjective experiments. PIXEL is seen to consistently outperform the other methods, while PATCH is often the second. PARAM consistently performs the worst. Rating questionnaires show that patches generated using PIXEL are also seen consistently as "plausible" fits for the reference regions; with results averaging at 1.0 in Figure 4(b) ('+'), while consistently residing above neutral (0), meaning negative ratings ('-' or '--') were almost nonexistent.

Post-hoc tests shown in Figure 4(d) were performed with Tukeys Honestly Significant Difference (HSD) method,[31] in order to determine which methods are statistically different from each other. All 3 studies indicated statistically significant difference of PIXEL from other methods and produced identical groupings among methods. Despite PATCH performing slightly better than OPTIM in results, they were not found to different at a statistically significant level. PARAM is found statistically significantly the poorest method of all four according to each questionnaire. All questionnaires consistently indicated the same Tukey's HSD grouping for the methods.

In Table 1, two-way repeated measures analysis of variance (rANOVA) was used to show that our main conclusion that PIXEL is seen to consistently outperform the other methods is significant. The p-value gives the probability of the null hypothesis being true. In our case, $p<0.01$ for all three questionnaires, which in combination with a high F-value shows that the null hypothesis can be rejected with high probability, regardless of the questionnaire technique used. On the other hand, the effect of participant expertise (i.e., the number of ultrasound scans performed) on the judgment of the individual methods did not lead to significant differences in judgment; potentially due to the superiority of PIXEL being obvious at any expertise level. Neither did other attributes like specialization. Thus, we can answer the two following questions confidently: First, PIXEL outperforms any other synthesis method. Second, the three questionnaires demonstrate consistent results, where the choice of questionnaire seems insignificant for the realism question studied; this can be interpreted that one could in the future use the simplest or quickest questionnaire variant for a user study.

4.2 Objective quality estimation

We next sought an automatic predictor to assess the realism of speckle appearance in an US image patch. Since speckles present properties of texture patterns, we investigated standard metrics used for texture and image assessment in computer vision and graphics described in Section 2. We first assessed correlation of individual features with assessed patch realism, and then also built a regression model upon them.

Table 1. rAnova results for significance of the superiority of PIXEL over other texture synthesis methods, the expertise level of the participants, and the specialization.

	PIXEL		Expertise		Specialization	
Questionnaire	F-value	p-value	F-value	p-value	F-value	p-value
Paired Comparison	81.3	1.4e-13	1.5e-31	1	4.3e-32	1
Rating	98.1	3.4e-15	0.9	0.47	0.85	0.47
Ranking	113.3	1.4e-16	4.2e-32	1	1.5e-31	1

First, we tested linear correlation (Pearson coefficients) of rating results with the texture features listed above. They range from -1 to 1, where -1 denotes negative, 1 denotes strong correlation between variables, and 0 no correlation. We tested the metrics induced by SSIM, the synthesizability features individually (LBP, SFilter, GIST, textureness, homogeneity, repetitiveness and irregularity), GLCM, Textons, and grayscale histogram comparisons using Bhattacharyya distances. The highest correlation coefficients were obtained with GIST (0.35), LBP (0.34), and textons (0.19). The remaining metrics (e.g., SSIM) presented less than marginal correlation. This shows that a direct correlation of such texture metrics with realism assessment is little to nonexistent. Computing the linear correlation with the combination of synthesizability features (all linearly combined as defined by Dai et al.[23]), we achieved a minor improvement (0.44).

Second, we trained a regression model using machine learning to predict the rating of a patch given its features – basically modeling the human rater. For observations (and later for realism predictions) the 44 mean Likert scale ratings of all synthesized textures were used. We employed a regression forest of 50 trees using least-squares boosting, with leave-one-patch-out experiments, where the rating of one patch was predicted given the remaining 43 ratings. The correlation of the resulting 44 estimates to their user-assessed mean values were then used to evaluate predictor performance. We applied an empirical feature selection process by selecting multiple combinations of features. Using textureness, repetitiveness, GIST, LBP distances (both standard and rotationally invariant), texton distances (both standard and rotationally invariant), and Bhattacharyya distances altogether led to the maximum correlation coefficients of 0.64 Pearson and 0.66 Spearman.

This result shows that machine learning is viable as an automatic predictor for perceived ultrasound texture similarity, and could be used as objective quality metric for future works on texture synthesis.

5. CONCLUSIONS AND FUTURE WORK

Our study investigates the example-based synthesis of locally *homogeneous* US image regions, in particular US speckle appearance. Our paper shows that the fine details and variations in US speckle of actual in-vivo images can be replicated by certain texture synthesis methods. Our results can be useful for various simulation tasks, such as inpainting arbitrary anatomical models with realistic speckle and generating rare pathological cases.
The results show that pixel-based texture synthesis is the superior choice for creating plausible ultrasound speckle. A possible reason is that the other methods exploit patch coherence, which may create noticeable repetitions – desirable for man-made texture but less suitable for randomly distributed US speckle pattern. In more time-critical contexts where some visual errors are acceptable, a user may consider PATCH (i.e., texture quilting) as a faster alternative, since PIXEL is relatively slower due to its per-pixel nature.
Our regression forest results show that a predictor from texture features appears to be feasible. Further investigation into other features and machine learning techniques are planned in the future to further explore this direction.

Despite the PIXEL method clearly outperforming other methods, it is not clear yet if it is a sufficiently good method: rating results are satisfactory, but not perfect. It is also possible that users might have rated relatively to other presented methods, even in the case of Rating test, rather than in an absolute manner. Investigating this phenomenon and more sophisticated texture synthesis techniques, e.g. the recent deep learning-based ones,[13] is a potential future work. Moreover, it may be possible to add application-specific methods upon texture synthesis (e.g., global feature maps) to allow for spatial variations.[29] Finally, correlation is significant, but there is still margin for improvement. A detailed analysis of perceptually motivated metrics combined with machine learning could form a better US speckle quality predictor.

Acknowledgment. The authors would like to thank the Department of Obstetrics, Director Prof. Roland Zimmermann, and the Department of Gynecology, Director Prof. Daniel Fink, at the University Hospital Zurich for their participation in our survey. Funding was provided by the Swiss Commission of Technology and Innovation (CTI). Orcun Goksel acknowledges the support of the Swiss National Science Foundation. Kenneth Vanhoey acknowledges the support of the ERC grant "VarCity" (#273940).

REFERENCES

1. Goksel, O., Salcudean, S.E.: B-mode ultrasound image simulation in deformable 3-d medium. IEEE Transactions on Medical Imaging **28**(11) (Nov 2009) 1657–1669
2. Reis, G., Lappe, B., Kohn, S., C.Weber, Bertram, M., Hagen, H.: Towards a virtual echocardiographic tutoring system. In: Visualization in Medicine and Life Sciences. (2008) 99–119
3. Mattausch, O., Goksel, O.: Monte-carlo ray tracing for realistic ultrasound training simulation. In: Eurographics Workshop on Visual Computing for Biology and Medicine, Springer (October 2016)
4. Wei, L.Y., Lefebvre, S., Kwatra, V., Turk, G.: State of the art in example-based texture synthesis. In: Eurographics 2009, State of the Art Report, EG-STAR, Eurographics Association (2009)
5. Bamber, J.C., Dickinson, R.J.: Ultrasonic b-scanning: a computer simulation. Physics in Medicine and Biology **25**(3) (1980) 463
6. Meunier, J., Bertrand, M., Mailloux, G.: A model for dynamic texture analysis in two-dimensional echocardiograms of the myocardium. SPIE **0768** (1987) 193–200
7. Gao, H., Choi, H., Claus, P., Boonen, S., Jaecques, S., Van Lenthe, G., Van der Perre, G., Lauriks, W., D'hooge, J.: A fast convolution-based methodology to simulate 2-d/3-d cardiac ultrasound images. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **56**(2) (2009) 404–409
8. Chan, W.Y., Qin, J., Chui, Y.P., Heng, P.A.: A serious game for learning ultrasound-guided needle placement skills. IEEE Transactions on Information Technology in Biomedicine **16**(6) (Nov 2012) 1032–1042
9. Mattausch, O., Goksel, O.: Scatterer reconstruction and parametrization of homogeneous tissue for ultrasound image simulation. In: IEEE Engineering in Medicine and Biology Conference (EMBC), EMBC (August 2015)
10. Onofrey, J.A., Oksuz, I., Sarkar, S., Venkataraman, R., Staib, L.H., Papademetris, X.: Mri-trus image synthesis with application to image-guided prostate intervention. In: MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), Springer International Publishing (2016) 157–166
11. Heeger, D.J., Bergen, J.R.: Pyramid-based texture analysis/synthesis. In: Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '95, ACM (1995) 229–238
12. Portilla, J., Simoncelli, E.P.: A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision **40**(1) (October 2000) 49–70
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems 28. (May 2015)
14. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of Conference on Computer Vision. ICCV '99, IEEE Computer Society (1999) 1033–
15. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Proceedings of Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '01, ACM (2001) 327–340
16. Ashikhmin, M.: Synthesizing natural textures. In: Proceedings of the 2001 Symposium on Interactive 3D Graphics. I3D '01, ACM (2001) 217–226
17. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '01, ACM (2001) 341–346
18. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. ACM Transactions on Graphics, SIGGRAPH 2005 (August 2005)
19. Daly, S.: The visible differences predictor: an algorithm for the assessment of image fidelity. In Andrew B. Watson, ed.: Digital images and human vision. MIT Press, Cambridge (oct 1993) 179–206
20. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4) (2004) 600–612
21. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision **62**(1) (2005) 61–81
22. Okura, F., Vanhoey, K., Bousseau, A., Efros, A.A., Drettakis, G.: Unifying color and texture transfer for predictive appearance manipulation. Computer Graphics Forum (2015)
23. Dai, D., Riemenschneider, H., Van Gool, L.: The synthesizability of texture examples. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)

24. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7) (July 2002) 971–987
25. Schmid, C.: Constructing models for content-based image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 2. (2001) 39–45
26. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42**(3) (2001) 145–175
27. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**(6) (November 1973) 610–621
28. Fukunaga, K.: Introduction to Statistical Pattern Recognition (2Nd Ed.). Academic Press Professional, Inc., San Diego, CA, USA (1990)
29. Kaspar, A., Neubert, B., Lischinski, D., Pauly, M., Kopf, J.: Self tuning texture optimization. Computer Graphics Forum **34**(2) (May 2015) 349–359
30. Mantiuk, R.K., Tomaszewska, A., Mantiuk, R.: Comparison of Four Subjective Methods for Image Quality Assessment. Computer Graphics Forum **31**(8) (dec 2012) 2478–2491
31. Tukey, J.W.: Comparing individual means in the analysis of variance. Biometrics **5**(2) (1949) 99–114