PERFORMANCE OF TWO MULTISCALE TEXTURE ALGORITHMS IN CLASSIFYING SILVER GELATIN PAPER VIA K-NEAREST NEIGHBORS

Kirsten R. Basinet[*], *Andrew G. Klein*[*], *Patrice Abry*[†], *Stéphane Roux*[†], *Herwig Wendt*[§], *Paul Messier*[‡]

[*] Engineering and Design / Computer Science, Western Washington Univ., Bellingham, WA 98225
[†] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France
[§] IRIT, CNRS UMR 5505, University of Toulouse, France
[‡] IPCH Lens Media Lab, Yale University, West Haven, CT 06516
contact author: andy.klein@wwu.edu

ABSTRACT

As part of the Historic Photographic Paper Classification Challenge, a multitude of approaches to quantifying paper texture similarity have been developed. These approaches have yielded encouraging results when applied to very controlled datasets containing photomicrographs of familiar specimens. In this paper, we report on the k-nearest neighbors classification performance of two multiscale analysis-based texture similarity approaches when applied to a much larger reference collection of silver gelatin photographic papers. The clusters for this data set were derived from a visual sorting experiment conducted by art conservators and paper experts later extended through crowd-sourcing. The results show that these texture similarity approaches, when combined with a simple k-nearest neighbors classification algorithm, yield workable performances with accuracy of up to 69%. We discuss this outcome in the context of available data and the cross-validation procedure used, then provide suggestions for improvement.

Index Terms— Texture similarity, photographic paper, crowdsourcing, multiscale analysis

1. INTRODUCTION

Texture is an essential characteristic of any photographic print, and paper idiosyncrasies can help facilitate the functional and expressive intentions of artists. After over 100 years of silver gelatin (traditional black and white) photographic paper manufacture, the profusion of textures can seem like a nearly infinite universe, defying all but the most basic attempts at visual classification.

Texture analysis provides important insights to the community of art scholars at museums and other collecting institutions. Understanding how a particular photographic paper was manufactured can help validate authenticity, identify purpose, and make important connections in the history of an artist or group of artists that may have worked together [1, 2]. Paper classification has traditionally been based on visual inspection by art conservators and curators [2]; however, several groups of researchers have begun researching alternate methods as part of the Historic Photographic Paper Classification Challenge [1, 3] and have developed various computational measures of texture similarity in the process [1, 2, 4–8]. These methods have demonstrated great promise on very controlled data sets, though their utility as a similarity measure when applied to larger, real-world data sets remains an open question.

Work supported by Grant ANR-16-CE33-0020 MultiFracs.

The Yale Lens Media Lab (LML) Reference Collection of Photographic Papers is perhaps the largest of its kind in the world and contains thousands of samples [9] from 65 manufacturers and more than 360 brands, serving as an invaluable resource for developing texture similarity approaches. The data set includes over 2,000 photomicrographs taken using magnification and raking light [1, 10] with each photopaper sample. Recently, a crowdsourcing experiment was conducted to classify the images in the LML reference collection into one of 6 groupings identified by art conservators and paper experts [11]. In this paper, we assess the performance of two previously-studied texture similarity quantifiers when used to classify images using a simple k-nearest neighbors algorithm. The two image processing techniques are both based on multiscale analysis; with one utilizing anisotropic wavelets [2], and the other fractals [4]. In addition to presenting the k-NN model cross-validation procedure and classification performance on these two approaches, we present the resulting confusion matrices and discuss the implications with respect to the algorithms' ability to help correctly – or incorrectly – categorize texture images. The results of this paper serve to validate the use of multiscale analysis approaches for assessing texture similarity in large, real-world databases.

2. DESCRIPTION OF DATA SET

In 2007, an experiment conducted at the Museum of Modern Art made a pioneering attempt to identify major texture groupings by tasking 19 domain experts with sorting 81 texture images chosen from the LML reference collection. The 6 texture categories ultimately established by the group were analyzed through hierarchical clustering, and showed that the observers largely shared agreement across 6 "protean" texture clusters, described in more detail in [12].

As it was infeasible for 19 professionals to manually cluster all 2,000 images in the LML reference collection, a crowd-sourced classification task was performed using workers on the Amazon Mechanical Turk (mTurk) platform [13]. The final classification labels assigned to each image were determined by simple majority consensus and ultimately demonstrated 92% agreement between the crowd and domain experts. In this expanded crowd-sourced experiment, 80 unique mTurk workers completed 130 person-hours over two days, with 90% of the work done by 23 mTurk workers. The crowd-sourced classification resulted in a rich set of data, with each sample receiving 24 "votes" which provided a distribution of the likelihood that each image was a member of each group. The reader is encouraged to consult [11] for detailed information about the crowd-sourced classification process.

3. TEXTURE CHARACTERIZATION TOOLS

As they were fully described elsewhere [2,4], we only provide here a qualitative description of the two texture processing tools considered herein, emphasizing features and distances on which they rely.

3.1. Anisotropic Multiscale Analysis (AMA)

Anisotropic multiscale analysis (AMA) [14] has been proposed in the context of the analysis of scale-free (or scale invariant) textures. It relies on the use of the Hyperbolic Wavelet Transform (HWT) [15]. The HWT consists of a variation of the 2D-Discrete Wavelet Transform (2D-DWT) [16], that explicitly takes into account the possible anisotropic nature of image textures. Indeed, instead of relying on a single dilation factor a used along both directions of the image (as is the case for the 2D-DWT), HWT relies on the use of two independent factors $a_1 = 2^{j_1}$ and $a_2 = 2^{j_2}$ along directions the horizontal (x_1) and vertical (x_2) directions. The HWT coefficients of imaged paper i are defined as inner products against wavelet templates, dilated with horizontal and vertical factors $a_1; a_2$ and translated at location k_1, k_2: $T_i((a_1, a_2), (k_1, k_2)) = \langle i(x_1, x_2), \frac{1}{\sqrt{a_1 a_2}}\psi(\frac{x_1-k_1}{a_1}, \frac{x_2-k_2}{a_2})\rangle$. Structure functions, consisting of space averages of the $T_i((a_1, a_2), (k_1, k_2))$ at scales a_1, a_2, are computed: $S_i((a_1, a_2), q) = \frac{1}{n_a}\sum_{\underline{k}} |T_i((a_1, a_2), (k_1, k_2))|^q$, with n_a the number of $T_i((a_1, a_2), (k_1, k_2))$ actually computed. To ensure that features do not depend on image intensity and that all scales contribute to texture characterization, the features consist of log-transformed normalized structure functions $\tilde{S}_i(a, q) = \ln \frac{S_i(a,q)}{\sum_{a'} S_i(a',q)}$. We use here $q = 2$ and a vector of 7 dyadic scales $a = 2^l$, ranging from 2 pixels (6.51μm) to 2^7 (834μm), for a total of $7 \times 7 = 49$ features $\tilde{S}_i(a, q)$.

To measure proximity between two images i and j, a L^p norm cepstral-like distance is computed (here we use $P = 1$):

$$D(i, j) = \left(\sum_a |\tilde{S}_i(a, q) - \tilde{S}_j(a, q)|^p\right)^{\frac{1}{p}}.$$

3.2. Pseudo-area-scale analysis (PASA)

The PASA approach [4] uses fractal analysis to decompose a surface into a patchwork of triangles of a given size. As the size of the triangles is increased, smaller surface features become less resolvable and the 'relative area' of the surface decreases. The topological similarity of two surfaces is computed by comparing relative areas at various scales. Though photomicrographs taken using magnification and raking light do not provide a direct measure for height, light intensity (i.e., pixel brightness) is used as a proxy for height.

PASA first extracts a square $N \times N$ region from the center of the image (where N was chosen to be 1024), and normalizes the intensity of the resulting extracted image. The grid of N^2 equally spaced points (representing pixel locations) is decomposed into a patchwork of $2(\frac{N-1}{s})^2$ isosceles right triangles where s is a scale parameter representing the length of two legs of each triangle. The pixel values at each of the triangle vertices are then taken as the 'pseudo-height' of each of the vertices. The area of each triangle in 3-D space is then computed and the areas of all triangular regions are summed, resulting in the total relative area A_s at the chosen scale s, serving as features. a vector $\mathcal{S}$ of scales s ranging from 1 pixel to 34 pixels, (6.51 μm to 0.221 mm), for a total of 8 features. To assess the similarity of two images i and j, a χ^2 distance measure $d(i, j)$ is computed via

$$D(i, j) = \sum_{s\in\mathcal{S}} \frac{\left(A_s^{(i)} - A_s^{(j)}\right)^2}{A_s^{(i)} + A_s^{(j)}}.$$

where $A_s^{(i)}$ is the vector of relative areas, used as features, computed for To conduct feature extraction, the relative area for an image is computed over a range of scale values; in this study, 8 scale values were used ranging from 1 pixel to 34 pixels, which correspond to lengths of 6.51 μm to 0.221 mm, respectively.

Small values of $d(i, j)$ indicate high similarity between images i and j, while large values indicate low similarity.

4. AUTOMATED CLASSIFICATION VIA K-NEAREST NEIGHBORS ALGORITHM

4.1. k-Nearest Neighbors Algorithm

k-nearest neighbors (k-NN) is a simple and well-known machine learning algorithm that uses some "nearest" group of training data around a test data point to classify it. [17]

For this application, we treat the AMA and PASA algorithms each as a distance function that k-NN uses to classify photomicrographs in texture categories 1-6. Subsequently, we evaluate the accuracy to gain insight into the usefulness of texture "distance", as calculated by AMA and PASA, as features for classifying paper samples.

In the simplest k-NN implementations, a test point is classified simply by assigning it to the most common class among the k nearest training points. In the interest of building a "fairer" classifier, we test this unweighted approach against both rank- and distance-weighted k-NN variations as described in [18]. This is achieved by calculating the contribution of each texture category by summing the inverse distances of the corresponding training samples (in the distance-weighted method) or the inverse rank of their similarity to the test sample (in the rank-weighted method) and accepting the category with the largest sum. In the event that multiple categories have the same value, k is decremented until the tie is broken.

4.2. Dataset partitioning and cross-validation

The dataset was partitioned into training/validation and test sets to isolate the model-building and evaluation stags. We used the original 64 images classified by domain experts as the test set, and this same test set was employed in [11]. To create the training and validation set, we first excluded images from the mTurk-classified set of 2,000 where the crowd's consensus was less than or equal to 50%. This was done to restrict attention to images where there was large agreement that the image was a member of a particular group, and reduced the training/validation set to 1,413 images. Some groups contained many more images than others as a result, with the largest group containing 490 images and the smallest group having 120.

To optimize the classifier models' two hyperparameters (the value of k in k-NN and a weighting method out of 'unweighted', 'rank', and 'distance' as described in Section 4.1), we used the popular "k-fold cross validation" technique. We clarify that the k in the name of this cross-validation technique is distinct from the k in the k-NN algorithm, and in this paragraph only, k refers to the number of folds. Under this approach, we randomly partitioned the original 1,413 images into $k = 10$ "folds" or equal sized subsamples, and iteratively used each of the k subsamples as cross-validation data for testing the model. Meanwhile, the remaining $k - 1$ subsamples were used as training data so that each of the k subsamples was used exactly once as cross validation data. We selected $k = 10$ because

of the reported optimal tradeoff of reliability and efficiency for that number of folds [19].

We then performed 5 randomized trials (by shuffling the dataset according to random seeds) of each 10-fold cross-validation for both AMA and PASA with all possible combinations of the hyperparameters. The final models were selected by observing the weightings method and values of k in the k-NN algorithms for AMA and PASA that resulted in the largest Top-1 accuracy. Note that **Top-1** accuracy is simply the proportion of times that the model prediction agrees with the mTurk crowd-assigned label, though we also report on **Top-2** accuracy (i.e., the proportion of times that the model's top 2 predictions agree with the true label).

Finally, we note that when the aforementioned hand-sorting experiment was conducted by the domain-experts, Group 1 was a somewhat miscellaneous group that included a variety of highly stippled textures that were distinct from the other textures in the group. As such, we decided to move these stippled images into their own "seventh" group, and the k-NN classifier was built to classify into 7 groups instead of the original 6. Subsequently, at the output of the k-NN classifier, any elements which were deemed to be part of this new seventh group were reassigned to be classified into Group 1, and so all results reported in the next section evaluate the ability of the k-NN classifier to classify into the original 6 groups. The choice between using 6 and 7 groups was in fact another hyperparameter selected during cross-validation, but we will not discuss this at length in the interest of presenting uncluttered results.

5. RESULTS

5.1. Cross-validation results

In order to select an optimal k and weighting scheme for the k-NN models on AMA and PASA, a 10-fold cross-validation procedure was performed on a training set of 1,413 images as described in Section 4.2. The best average Top-1 and Top-2 accuracy metrics found across values of k from 1 to 1,271 (the size of 9 training folds) in 5 randomized trials of cross-validation are given in Table 1 for both AMA and PASA. In computing the cross-validation results, 5 trials were deemed sufficient because the standard deviation of the average Top-1 and Top-2 accuracy for each k was no greater than 0.34% for any iteration of the test, meaning that we would not expect significantly different accuracy results from more trials.

Weighting	Best Avg. Top-1	Best Avg. Top-2
AMA		
None	75.17% at $k = 9$	95.06% at $k = 25$
Rank	78.36% at $k = 19$	95.78% at $k = 43$
Distance	74.73% at $k = 3$	94.73% at $k = 25$
PASA		
None	71.59% at $k = 5$	89.84% at $k = 9$
Rank	72.96% at $k = 15$	90.80% at $k = 11$
Distance	68.87% at $k = 5$	88.93% at $k = 13$

Table 1: Best average accuracies of k-NN weighting schemes for AMA and PASA over 5 randomized trials of 10-fold cross-validation

As shown in Table 1, the best possible Top-1 accuracy was achieved during cross-validation using rank weighting at $k = 19$ and $k = 15$ respectively for AMA and PASA. The average accuracy for both algorithms as a function of k is visualized in Fig. 1, and suggests that in general the choice of k does not have a significant impact on accuracy. (Data for $k > 60$ is not shown but was calculated.) This finding is a logical outcome of the rank-weighted scheme, since each additional neighbor included in the classification calculation will contribute increasingly little to the final prediction.

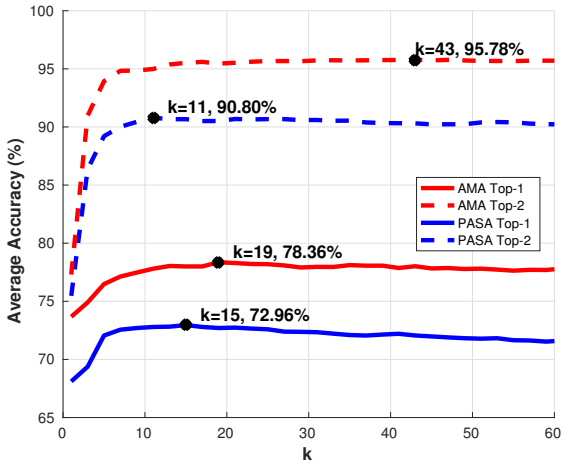


Fig. 1: Average accuracy of k-NN on AMA and PASA using rank weighting during 5 randomized trials of 10-fold cross-validation. Global maxima indicated with $*$.

5.2. Test results

From the process of cross-validation on the training set of 1,413 images, we built two rank-weighted k-NN models and found the optimal value of k to be 19 for AMA and 15 for PASA as described in Section 5.1. In this section, we report the results of running the optimal models on the test set of 64 images classified by domain experts.

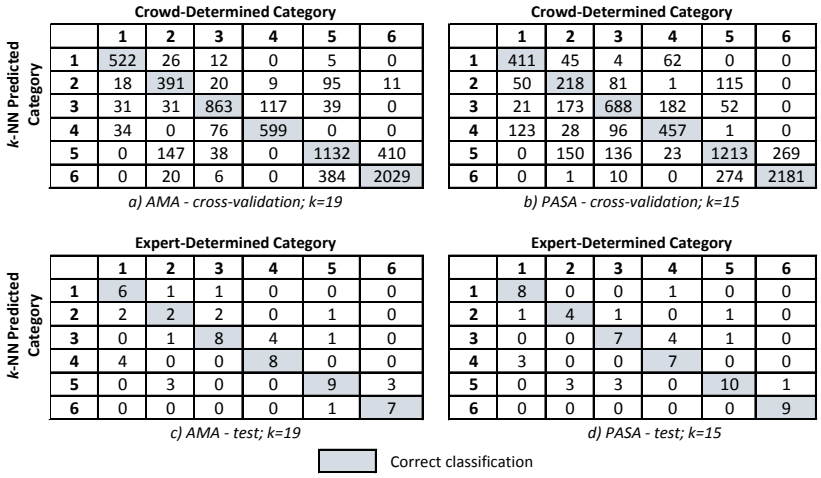
Because the distribution of categories was not uniform across the dataset, it is important to consider the results in relation to the size of largest class: for the dataset of 1,477 images used in this paper, class 6 held the majority with 500 images or about 34% of the total. Therefore, the Top-1 accuracy is only meaningful relative to that baseline, since a model that simply classified every image as class 6 would be 34% accurate.

The accuracy results for the constructed k-NN models for AMA and PASA using training and test sets of 1,413 and 64 images respectively are displayed in Table 2, showing that k-NN applied to PASA outperforms AMA in Top-1 accuracy but vice-versa for Top-2 accuracy. Given the significant advantage of AMA in all cases during the cross-validation phase, it is surprising that PASA's k-NN model is 9% more accurate in Top-1 classification.

	Top-1 Accuracy	Top-2 Accuracy
AMA, $k = 19$	62.50%	89.06%
PASA, $k = 15$	68.75%	87.50%

Table 2: Test results of k-NN on AMA and PASA

To help understand the behavior of k-NN on each algorithm, confusion matrices are provided in Fig. 2 and show the number of images from each "true" (crowd- or expert-determined) category that were predicted to be certain categories by k-NN. Results at the optimal k value for each algorithm are shown for the 5 trials of 10-fold



	1	2	3	4	5	6
1	522	26	12	0	5	0
2	18	391	20	9	95	11
3	31	31	863	117	39	0
4	34	0	76	599	0	0
5	0	147	38	0	1132	410
6	0	20	6	0	384	2029

a) AMA - cross-validation; k=19

	1	2	3	4	5	6
1	411	45	4	62	0	0
2	50	218	81	1	115	0
3	21	173	688	182	52	0
4	123	28	96	457	1	0
5	0	150	136	23	1213	269
6	0	1	10	0	274	2181

b) PASA - cross-validation; k=15

	1	2	3	4	5	6
1	6	1	1	0	0	0
2	2	2	2	0	1	0
3	0	1	8	4	1	0
4	4	0	0	8	0	0
5	0	3	0	0	9	3
6	0	0	0	0	1	7

c) AMA - test; k=19

	1	2	3	4	5	6
1	8	0	0	1	0	0
2	1	4	1	0	1	0
3	0	0	7	4	1	0
4	3	0	0	7	0	0
5	0	3	3	0	10	1
6	0	0	0	0	0	9

d) PASA - test; k=15

Fig. 2: Confusion matrices for k-NN models indicating the total correct and incorrect classifications in each category

cross-validation as well as the test dataset. Below we present some interesting observations from the matrices:

- The k-NN models for both AMA and PASA frequently mislabeled class 2 images as class 5, class 4 images as class 3, and class 6 images as class 5 in cross-validation and testing
- Class 2 was always the most commonly misclassified group
- k-NN with PASA was particularly good at identifying class 6 images ($\approx$90% accuracy) in both cross-validation and testing
- The classification performance on several categories changed significantly between the cross-validation and test stages, but in different ways for AMA and PASA (e.g. accuracy for class 2 decreased 35.01% in absolute terms with AMA and increased 21.73% with PASA; accuracy for class 1 decreased 36.28% for AMA but only 1.23% with PASA)
- The distribution of categories in the test set was much more uniform than in the training set

Although more context is required to determine the relationship between this information and the test results, the confusion matrices allow some speculation about the cause of inconsistencies between cross-validation and testing results. Large false-positive and false-negative entries in the matrices may suggest that there are aspects of AMA and PASA which are less ideal for comparing certain types of texture images. Indeed, the high Top-2 performance of both models (indicating that most texture images were either classified correctly or "just" missed) suggests that the algorithms are more intelligent than the Top-1 metric alone can convey; targeted improvements in processing certain texture class features may therefore measurably increase Top-1 accuracy. In addition, the dissimilar distributions underlying the training and test datasets are likely a source of unreliability; for instance, there is a larger proportion of class 2 images – the most error prone group – in the test set.

6. CONCLUSIONS AND PERSPECTIVES

In this paper, we described a k-nearest neighbors approach to classifying photographic paper textures by using calculations from two multiscale analysis-based texture comparison algorithms (AMA and PASA) in order to make a preliminary assessment of their usefulness with uncontrolled datasets. In the cross-validation phase with a training set of 1,413 images labeled by mTurk workers, we found that the k-NN models for AMA and PASA were able to achieve up to 78.36% and 72.96% Top-1 accuracy respectively. When testing on a domain expert-labeled dataset of 64 images, those accuracies fell to 62.50% and 68.75%, though Top-2 accuracies approached 90%.

Although some performance loss is to be expected when moving from a training to test scenario, we did not anticipate that the Top-1 accuracy of the k-NN model for AMA would fall below that of PASA. While a larger test set is needed to examine the cause, it is probable that dissimilar distributions and characteristics of images between the training and test sets are partly responsible. It also cannot be ignored that the classifications assigned to the training set by non-expert mTurk workers were used to evaluate the accuracy of test images classified by domain experts. While a previous study has shown that the mTurk workers most often classify images in consensus with domain experts [11], it is possible that some images used to train the k-NN models were labeled incorrectly (i.e., domain experts might disagree with the crowd in some cases).

Nonetheless, the initial results demonstrate the feasibility of using both AMA and PASA to measure texture similarity and provide features for classification, and a logical extension of the research would investigate other classification algorithms and include other available metadata for each texture sample (such as reflectance, paper brand, etc.) Further, this study demonstrates that such schemes not only share an internal logic, making for useful comparisons, but also bear a measurable relationship to human perception. Future work to isolate the specific strengths and weaknesses of computational methods alongside expert observation has the potential to refine the classification algorithms as well as help experts achieve heightened visual acuity and incisiveness. Continued work along these lines also holds promise of more precisely modeling sets of features that drive human perception and classification of textured surfaces. More practically, expert systems based on careful characterization and automated sorting of visual/tactile attributes could form the basis of networked platforms that would enable the discovery of material-based affinities among objects, in the fine art domain and beyond.

The data set of photographic paper textures used in this paper can be obtained by emailing the contact author.

7. REFERENCES

[1] C Richard Johnson, Paul Messier, William A Sethares, Andrew G. Klein, Christopher Brown, Anh Hoang Do, Philip Klausmeyer, Patrice Abry, Stephane Jaffard, Herwig Wendt, et al., "Pursuing automated classification of historic photographic papers from raking light images," *Journal of the American Institute for Conservation*, vol. 53, no. 3, pp. 159–170, 2014.

[2] Patrice Abry, Stephane G Roux, Herwig Wendt, Paul Messier, Andrew G Klein, Nicolas Tremblay, Pierre Borgnat, Stephane Jaffard, Beatrice Vedel, Jim Coddington, et al., "Multiscale anisotropic texture analysis and classification of photographic prints: Art scholarship meets image processing algorithms," *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 18–27, 2015.

[3] Paul Messier, "Paper Texture ID Challenge," 2013.

[4] Andrew G Klein, Anh H Do, Christopher A Brown, and Philip Klausmeyer, "Texture classification via area-scale analysis of raking light images," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, Nov. 2014, pp. 1114–1118.

[5] W. A. Sethares, A. Ingle, T. Krč, and S. Wood, "Eigentextures: An SVD approach to automated paper classification," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 1109–1113.

[6] D. Picard and I. Fijalkow, "Second order model deviations of local Gabor features for texture classification," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 917–920.

[7] Y. Zhai and D. L. Neuhoff, "Photographic paper classification via local radius index metric," in *IEEE Intl. Conf. on Image Processing (ICIP)*, Sept. 2015, pp. 1439–1443.

[8] A. Sangari and W. Sethares, "Paper texture classification via multi-scale restricted boltzman machines," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 482–486.

[9] Paul Messier, "Conservation of photographs & works on paper," 2013.

[10] Paul Messier and C Richard Johnson, "Automated surface texture classification of photographic print media," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, Nov. 2014, pp. 1105–1108.

[11] Andrew G Klein, Paul Messier, et al., "Deep learning classification of photographic paper based on clustering by domain experts," in *Proc. Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 139–143.

[12] Khemraj Emrith, *Perceptual dimensions for surface texture retrieval*, Ph.D. thesis, Heriot-Watt University, 2008.

[13] Amazon.com, "Amazon mechanical turk," 2016.

[14] Stephane G Roux, Marianne Clausel, Beatrice Vedel, Stephane Jaffard, and Patrice Abry, "Self-similar anisotropic texture analysis: The hyperbolic wavelet transform contribution," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4353–4363, 2013.

[15] R.A. DeVore, S.V. Konyagin, and V.N. Temlyakov, "Hyperbolic wavelet approximation," *Constructive Approximation*, vol. 14, pp. 1–26, 1998.

[16] Stéphane Mallat, *A wavelet tour of signal processing*, Academic press, 1999.

[17] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, January 1967.

[18] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 4, pp. 325–327, Apr. 1976.

[19] Ron Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, San Francisco, CA, USA, 1995, IJCAI'95, pp. 1137–1143, Morgan Kaufmann Publishers Inc.