Deep Learning Classification of Photographic Paper Based on Clustering by Domain Experts

Andrew G. Klein*, Paul Messier†, Andrea L. Frost*, David Palzer*, and Sally L. Wood‡

*Western Washington Univ.
Bellingham, WA 98225
Contact email: andy.klein@wwu.edu

†IPCH Lens Media Lab
Yale University
West Haven, CT 06516

‡Electrical Engineering Dept.
Santa Clara University
Santa Clara, CA 95053

***Abstract*—Prior work on texture analysis of historic, photographic papers has focused primarily on measures of texture similarity. However, automated grouping or clustering of photographic paper textures in a way that is meaningful to art conservators remains an open problem. In this work a deep learning approach to automated classification is presented, for clusters derived from a human sorting experiment conducted by 19 art conservators and paper experts and subsequently extended through crowd-sourcing. The proposed approach uses a deep convolutional neural network, and results are presented on the performance in automatically classifying images when compared to human experts.**

I. Introduction

Texture is an essential attribute of a photographic print, relating to the functional and expressive intentions of the photographer, and manufacturers went to great lengths to produce a variety of textures for photographic paper. After over 100 years of silver gelatin (traditional black and white) photographic paper manufacture, the profusion of textures can seem like a bewildering and nearly infinite universe, defying all but the most basic attempts at visual classification.

Texture analysis of photographic paper provides important insights to the community of art investigators at museums and other art 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 set of artists that may have worked together [1], [2]. Classification of photographic paper has traditionally been based primarily on human inspection conducted by art conservators and museum curators [2]. However, several groups of researchers have begun addressing the issue of photopaper texture classification starting with the Historic Photographic Paper Classification Challenge [1], [3]. These prior works have focused primarily on measures of texture similarity, and include such approaches as multi-scale analysis (using anisotropic wavelets [2] or fractals [4]), non-semantic feature extraction (eigentextures [5], random-feature textons [1], deviation of local Gabor features [6]), local radius index [7], and restricted Boltzmann machines [8]. Those methods have demonstrated great promise while at the same time uncovering important questions for future study.

Advances in machine learning have raised the prospect of automated classification of photographic papers in which the learning algorithm implicitly develops the classification features. It may be used not only to reinforce the classifications of human experts, but also to perhaps identify human classification errors. In this paper we explore the application of deep learning to automatically classify photographic papers, using six clusters derived from an experiment involving 19 art historians and conservators who were asked to cluster 81 different photographic paper textures. Since a data set consisting of only 81 papers is insufficient for training deep neural networks, we explored the use of crowd-sourcing in the classification of photographic papers to generate a larger data set. An initial crowd-sourcing experiment was conducted on the 81 papers previously classified by domain experts, and the classification accuracy of the crowd was shown to be 92%. Subsequently, a crowd-sourcing experiment to classify 2,000 images of silver gelatin photographic paper textures was conducted, and the resulting data set was used to train a convolutional neural network (CNN). Section II describes the human sorting and crowd-sourcing experiments which were used to create the dataset labels, Section III describes the deep CNN architecture, and Section IV provides the classification results of the CNN.

II. Description of Data Set

The Paul Messier Historic Photographic Papers Collection contains thousands of samples of photographic paper, and catalogs the manufacturer, brand, surface texture, and reflectance [9]. This collection is perhaps the largest photographic paper collection in the world with samples from 65 manufacturers and more than 360 brands. The data set for the research described in this paper consists of more than 2,000 photomicrographs taken using magnification and raking light [1], [10] of each photopaper sample, like the examples shown in Fig. 1. Here, we describe how each image in the dataset was labeled/classified into one of six groups.

A. Human clustering

Conceived by Michael Chantler of the Heriot-Watt University Texture Lab, an experiment conducted at the Museum of Modern Art (MoMA) in 2007 made a first attempt to identify texture groupings that were apparent to human observers. This work was conducted using printed images of textures made with a high resolution scanner depicting 1 cm^2 of the surface of a sheet of photographic paper. From the complete data set of

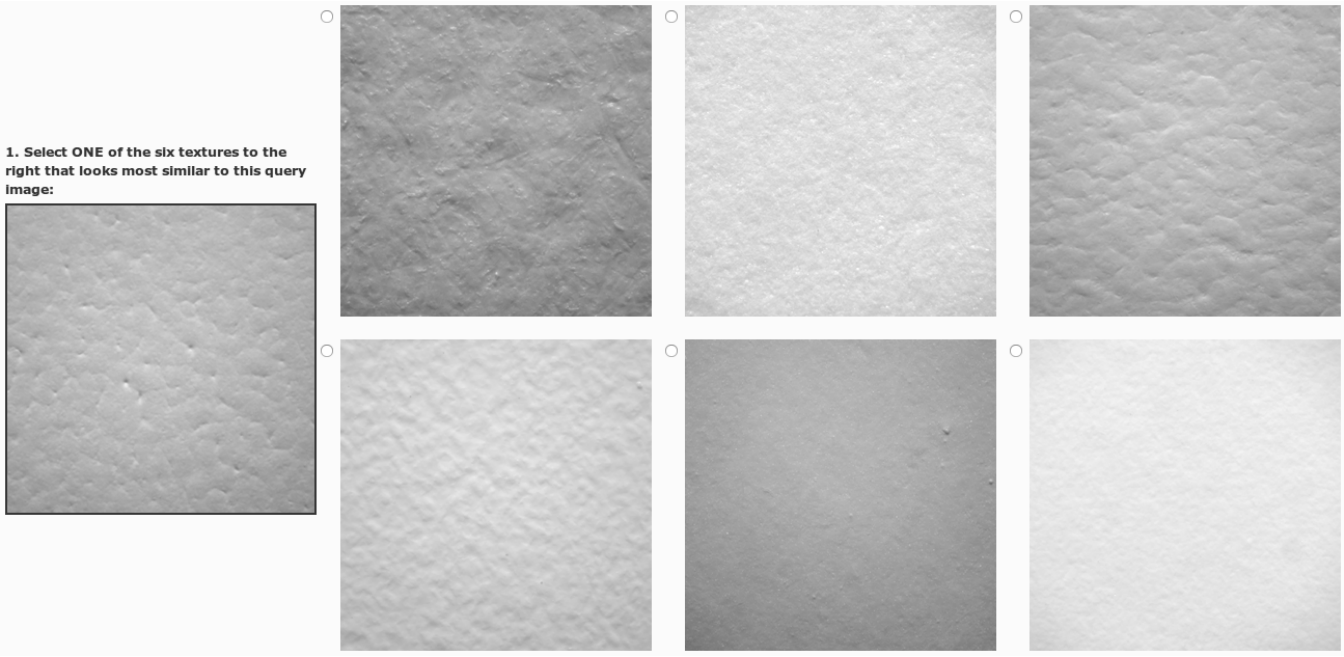

Fig. 1. Example question from crowd-sourcing experiment. Images along top row are random example textures from groups 1, 2, and 3; bottom row are from groups 4, 5, and 6. The query image is shown at the left, and in this case it happens to be a member of group 3.

2,000+ textures, a subset of 81 different papers were selected to represent the diversity of silver gelatin photographic paper textures. The task assigned to the human observers was to make any number of logical groupings from this subset. The groupings created by each of the 19 domain experts were analyzed through hierarchical clustering by K. Emrith, and showed the human observers largely shared agreement across six "protean" texture clusters which are described in more detail in [11].

This work provided a strong sense that the seemingly unordered and infinitely variable set of photographic paper textures could be objectively categorized into sets and subsets. Once identified, the common characteristics of these groupings can be recognized, gradually forming a shared visual vocabulary that would allow professionals, such as conservators and curators, to share and communicate information about photographic prints more effectively. With the advent of automated texture classification schemes [1], a logical next step was the development of a machine learning system able to recognize the same six clusters identified by the human observers. Once fully automated, a classification system built along these lines has the potential to sort extremely large collections of photographic paper, at volumes far exceeding the capacity of human experts and to discover affinities within and across geographically separated collections.

B. Crowd-sourced classification

Supervised machine learning requires a significant amount of training data in order to build a classifier that is able to generalize [12], and 81 images is considered to be a deficient amount of data for training a deep neural network. As it was infeasible for the 19 domain experts to manually cluster all 2,000 images, a crowd-sourced classification task was subsequently conducted using Amazon Mechanical Turk (mTurk) [13]. To verify the accuracy of crowds in classifying images into the 6 groups created by the domain experts, an initial experiment was conducted using the same 81 images that had been clustered by the domain experts. The mTurk workers were presented with multiple-choice questions consisting of a query image as well as 6 random images selected from each of the 6 clusters identified by the domain experts. The mTurk workers were then asked to identify which of the 6 images most resembled the query image; an example question is shown in Fig. 1. Each of the 81 textures was presented as the query image 24 times, and the 6 randomly selected images from each of the 6 groups varied each time. Since the 6 randomly selected images presented as choices were also drawn from the group of 81, each of the 81 images were also presented as possible choices many times. As with the human-sorting experiment, the images shown were 1 cm^2 of the surface of a sheet of photographic paper, though in this case they were illuminated by raking light. While the original grayscale images in the data set had a resolution of 1024×1024 pixels, they were downsampled by a factor of 4 in both dimensions, resulting in images of size 256×256 pixels. While higher resolution images might be preferable, this downsampling was performed so that the query image and all 6 choices would would fit on the majority of users' computer monitors.

Using a majority rule (where a given image was declared to be a member of the group garnering the most votes), the collective mTurk classifications demonstrated 70% agreement

with the domain experts, though the crowd's top 2 choices matched domain experts 95% of time. It was observed that many of the "misclassified" images exhibited characteristics from more than one of the 6 groups, and, as with the results on the human-sorting experiment, some images were consistently classified as being in two or more groups [11]. Thus, to arrive at a more distinct set of baseline images to present as *choices* in the crowd-sourced multiple-choice questions, we repeated the crowd-sourced replication of the human-sorting experiment by omitting 17 of the original 81 images where there was disagreement between mTurk workers and domain experts, leaving $81 - 17 = 64$ images, with a roughly equal number of images in each of the 6 groups. Upon repeating the crowd-sourced classification on the reduced set of 64 images, agreement between mTurk workers and domain experts rose to a top-1 accuracy 92%, and a top-2 accuracy of 100%.

Having established the efficacy of crowd-sourcing to replicate domain expert classification of photographic papers into the 6 groups with high accuracy, we proceeded to use mTurk to classify all 2,000 images in the data set, using the 64 human-classified images as choices. In this expanded crowd-sourced experiment, 5% of the questions consisted of "known answers" where the query image matched perfectly with one of the 6 choices (though often with horizontal flipping and/or brightness adjustment). These known answers were embedded in the crowd-sourcing task to measure the performance of the mTurk workers. Workers that consistently submitted results where the known answers were correct received small monetary bonuses; workers that submitted results where the known answers were incorrect were warned, and in some cases penalized for repeated submission of inaccurate work. The mTurk workers completed 130 person-hours over two days, and involved 80 unique participants, with 90% of the work done by 23 mTurk workers. The crowd-sourced classification resulted in a rich set of data for training a neural network, with each of the 2,000 having received 24 "votes" which provided a distribution of the likelihood that each image was a member of each group.

III. Automated Classification via CNN

A. Dataset partitioning

Before training the CNN, the dataset was first partitioned into training, validation, and test sets. We used the original 64 images classified by domain experts as the test set; under this partitioning, we implicitly used those 64 images to "train" the crowd to classify the remaining 2,000 images which were then used to train the neural network which was then tested on the original test set of 64 human-sorted images. This circular idea is shown in the Fig. 2.

To create the training and validation sets, we first excluded images from the 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 data set to 1,413 images. Some groups had far more images than others, however, with the smallest group having only 120 images.

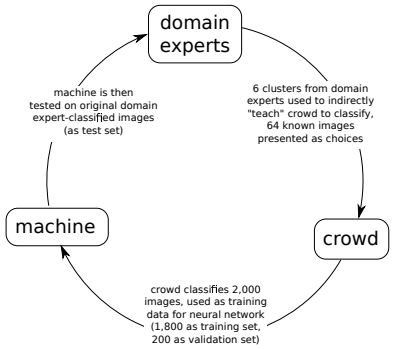

Fig. 2. Interaction of domain experts, crowd classification, and machine classifier.

In classification problems, it is generally preferable to have a roughly equal number of examples of each classification. As such, we randomly selected 120 images from each group, resulting in a subset of $120 \times 6 = 720$ total images. These 720 images were then randomly partitioned into training and validation sets in the fraction of 85% and 15%, respectively. Thus, we had distinct images in each partition, with 64 images in the test set, 612 in the training set, and 108 in the validation set. Again, all images were 1024×1024 pixel grayscale images.

B. Data normalization, tiling, and augmentation

An interesting research question concerns the resolution (or physical scale) necessary to perform accurate automated classification of photographic paper textures. The grayscale images in the data set had a resolution of 1024×1024 pixels representing 1 cm^2 of physical area, though as mentioned previously the crowds were shown downsampled images of size 256×256 pixels. While larger resolution images may be preferable for more accurate representation of small texture features, larger resolution images place increasing demands on the required memory and computation to train the CNN. We varied the resolution of the input images to test the effect of classification performance[1], and found that images downsampled by as much as a factor of 4, to 256×256, showed no noticeable degradation in classification performance compared to full resolution images which suggested that the small features in the full resolution image may not be relevant to the present classification task. Indeed, the crowds were able to perform classification of the test set at this same lower resolution with 92% accuracy, and the labels of the training and validation sets were determined by the crowd at this lower resolution. Thus, we chose to provide the machine with the same downsampled 256×256 images, consistent with what we had provided the crowds.

To mitigate the effect of differing brightness and contrast in the grayscale images, we normalized them by subtracting

[1]Since neural networks with vastly different input sizes also require differing amounts of hidden nodes, it is challenging to conduct and apples-to-apples comparison of the effect of resolution. In our experimentation, we scaled the number of hidden nodes with the number of input size.

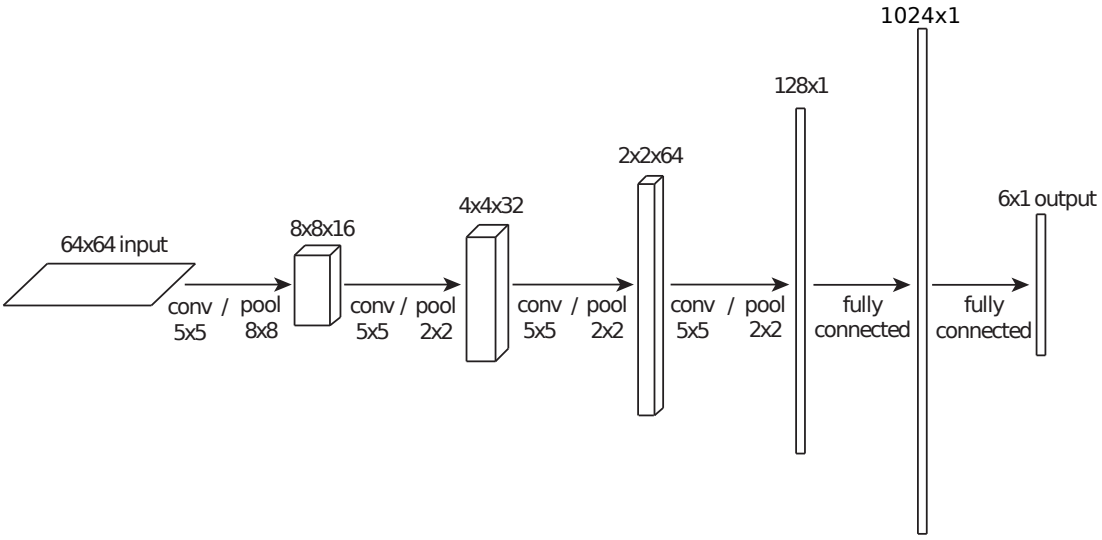

Fig. 4. Architecture of six-layer convolutional neural network.

the mean and then dividing by the magnitude, which resulted in a 256×256 matrix of values between ± 1. Then, to reap additional computational savings, we performed a 4×4 tiling of the 256×256 matrices, resulting in 16 tiles each of size 64×64, and we used these 64×64 tiles as the input to our neural network. To increase the effective size of our training and validation data, we used data augmentation [12] by taking an additional 3×3 tiling offset by 32 pixels on each side, and subsequently horizontally flipping the result, as shown in Fig. 3. We note that horizontal flipping preserves the direction

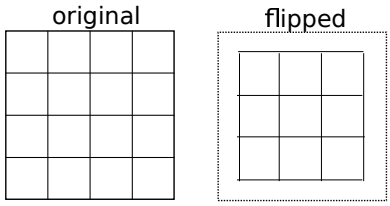

Fig. 3. Data augmentation via horizontal flipping and offset tiling to arrive at 25 total tiles per image. Each tile is 64×64 pixels, and the whole image is 256×256 pixels.

of the raking light used to create the silver gelatin photomicrographs. With the 16 tiles from the source image, plus 9 additional tiles from this label-preserving data augmentation step, each image was represented by 25 tiles. As described above in section III-A, the training and validation set contained 720 total images, with 85% of them used for training. Thus, the training set consisted of $720 \times 25 \times 85\% = 15,300$ tiles of size 64×64, which we hoped was a sufficient amount of training data to avoid overfitting.

C. Convolutional neural network architecture and training

The architecture for the neural network is shown in Fig. 4. We employed a neural network consisting of 6 total layers, 4 of which were convolutional, and 2 of which were fully connected. We borrowed many of the basic ideas from the ImageNet classifier in [12], while the specific parameters are quite different. The input to the neural network is a normalized 64×64 grayscale tile, and the output is a 6×1 vector of softmax outputs indicating the likelihood that the tile is a member of each of the 6 groups. All layers used a rectified linear unit (ReLU) activation function, i.e., the half-wave rectifier $f(x) = \max(x, 0)$, and all layers used a bias term. The convolutional layers all used 5×5 patches with zero-padding; the first of the four convolutional layers was followed by 8×8 max pooling, while the subsequent three convolutional layers employed 2×2 max pooling [12]. The fully-connected layers all used L_2 regularization, the first fully-connected layer used 50% dropout during training to prevent overfitting [14], and the second fully-connected (readout) layer used softmax.

To train the CNN, we employed a mini-batch size of 32 tiles, and used an exponentially decaying learning rate. A momentum optimizer was used, and the chosen cost function for training was the cross-entropy between mTurk workers' "majority vote" and the model's prediction. As training proceeded, the performance was computed on the validation set. For each validation image, the 25 6-vectors output for each of the 25 tiles were averaged, and the result provided the likelihood that a given image was in each of the 6 groups. The specific model was selected when the top-1 accuracy on the validation set was lowest, which occurred when the top-1 accuracy on the validation set was 84.3% correct; in addition, the top-2 accuracy was 98.0%. The CNN was implemented in Google's TensorFlow, trained on an NVIDIA GTX 1080 GPU, and required about 120 seconds to train.

IV. Results and Conclusions

Since the crowd's performance in classifying the test set was 92%, and the machine's performance in classifying the validation set was 84%, the path outlined in Fig. 2 suggests that the machine's performance on the test set could expected

to be roughly equal to the product of these two, i.e. 77%. Indeed, the CNN classification performance on the 64 images in the test set was quite close to this, with **top-1 accuracy of 79.7% and top-2 accuracy 92.2%**. We note that, with 6 groups, a random guess would have accuracy equal to 16.7%.

The first convolutional layer in a CNN processes the raw pixel data, so the weights of this layer are generally accepted as being the most interpretable, and give some indication as to the features that are emphasized by the CNN. Figure 5 shows all 16 of the 5×5 kernels of the first convolutional layer. The raking light that illuminates the photographic papers enhances the peaks and valleys of the texture, so there are quite frequently shadows that appear. Many of the kernels in Fig. 5 resemble these shadows.

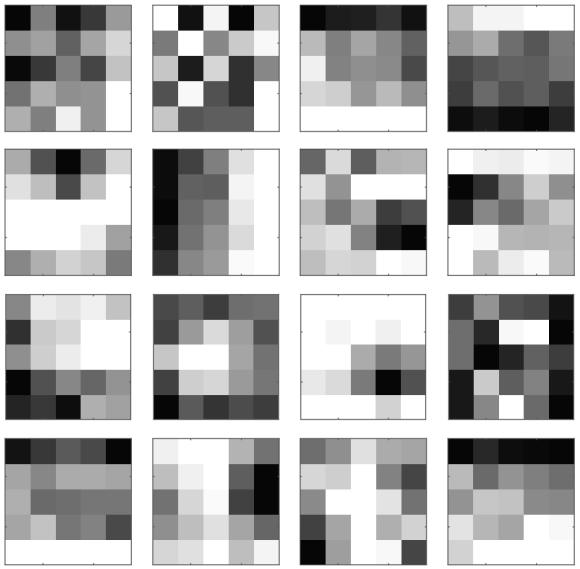

Fig. 5. Kernels of first convolutional layer.

Having achieved classification performance of nearly 80% on the test set, this work demonstrates the potential of deep learning for performing automated classification of photographic papers into groupings determined by domain experts. In addition, this work has shown the potential of using crowdsourcing for classifying photographic paper textures, and of using a relatively small sample set of images classified by domain experts to "train" crowds to classify a larger training set which ultimately facilitates use of supervised learning approaches.

In the present work, very little tuning of hyperparameters was conducted (e.g., adjusting the number of size of hidden layers, using alternate activation functions, optimizers, cost functions, etc). Future work could investigate exhaustive tuning of these CNN hyperparameters to achieve better classification performance. In addition, future work could investigate use of alternative machine learning structures that have recently shown promise in image classification, such as so-called "residual" neural networks [15].

Lastly, there are many other features of photographic paper that have not been previously considered for automated classification. For example, the 2,000 images in the photographic paper collection are accompanied by much additional metadata, including information about reflectance, brand, and surface texture descriptions, all of which could be used for automated classification.

All source code for this paper is available at [16]. In addition, the data set of photographic paper textures used in this paper can be obtained by emailing the contact author.

Acknowledgments

The authors would like to thank the many Amazon Mechanical Turk workers who painstakingly contributed to the crowd-sourced classification which was instrumental to this work.

References

[1] C. R. Johnson, P. Messier, W. A. Sethares, A. G. Klein, C. Brown, A. H. Do, P. Klausmeyer, P. Abry, S. Jaffard, H. 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] P. Abry, S. G. Roux, H. Wendt, P. Messier, A. G. Klein, N. Tremblay, P. Borgnat, S. Jaffard, B. Vedel, J. 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] P. Messier. (2013) Paper Texture ID Challenge. [Online]. Available: http://www.papertextureid.org/about.html

[4] A. G. Klein, A. H. Do, C. A. Brown, and P. 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. Kr, and S. Wood, "Eigentextures: An SVD approach to automated paper classification," in *Proc. Asilomar Conf. 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 *Proc. Asilomar Conf. 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 *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 1439–1443.

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

[9] P. Messier. (2013) Conservation of photographs & works on paper. [Online]. Available: http://www.paulmessier.com/#!collection/cbeq

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

[11] K. Emrith, "Perceptual dimensions for surface texture retrieval," Ph.D. dissertation, Heriot-Watt University, 2008. [Online]. Available: http://www.macs.hw.ac.uk/texturelab/files/publications/phds_mscs/KE/KhemPhDThesis.PDF

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] Amazon.com. (2016) Amazon mechanical turk. [Online]. Available: https://www.mturk.com/mturk/welcome

[14] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[16] A. G. Klein. (2016) Source code for classification of photographic paper using a convolutional neural network. [Online]. Available: https://github.com/agklein1/photopaper