Analysis, Attribution, and Authentication of Drawings with Convolutional Neural Networks

Steven J. Frank
Art Eye-D Associates LLC
steve@art-eye-d.com
https://orcid.org/0000-0002-7905-7792

Andrea M. Frank
Art Eye-D Associates LLC

Research Article

Keywords: convolutional neural network, attribution, authentication, drawings

Posted Date: January 30th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2520519/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at International Journal of Arts and Technology on January 23rd, 2023. See the published version at https://doi.org/10.1504/IJART.2022.10050958.
Abstract

We propose an innovative framework for assessing the probability that a subject drawing is the work of a particular artist. While numerous efforts have applied neural networks to classify two-dimensional works of art by style and author, these efforts — with few exceptions — have been limited to paintings. Drawings, which can involve multiple media with very different visual characteristics and greater susceptibility to damage than paint, present a more formidable challenge. Our technique is robust to the age and wear of a drawing as well as the possibility that it contains marks made with multiple drawing media. We obtained classification accuracies exceeding 90% using a five-layer convolutional neural network (CNN), which we trained on a curated set of drawing images attributed to Raffaello Sanzio da Urbino (1483–1520), known as Raphael, as well as drawings by his admirers, imitators, and forgers.

1. Introduction

Artificial intelligence techniques have attained increasing acceptance in the production of artwork and its analysis. “Deep learning” and other approaches have been used to categorize art images by movement and style, to classify paintings by artist rather than style, and to extract and analyze artists’ brushstrokes [1–10]. More recent efforts have extended this approach to study questions of authentication and attribution, including the possibility that multiple authors contributed to a painting [11, 12].

Few deep learning efforts have focused on drawings. A recent overview article [13] lists a single study involving analysis of drawings using artificial intelligence [14]. This is not surprising in view of the technical challenges that drawings pose. An artist can choose among many drawing media and may use more than one in a particular work. Some of these media are more friable and tend to smudge and smear more than others [15]. Different drawing tools produce visually distinct marks, and an artist’s characteristic style with one tool may not be — and generally will not be — visually similar across other media. As works primarily on paper supports, drawings are especially vulnerable to the effects of handling and substrate chemistry, which may add to already-existing physical flaws such as tears and stains. Collectors who would not dare deface the surface of a painting often added stamps and catalog entries to Old Master drawings. Such casual handling reflects the often utilitarian purposes of a drawing. Works that are highly prized today may have been produced as preparatory sketches or cartoons for paintings, as plans for statues or architecture, or as finger exercises. While experts have firmly attributed fewer than 20 paintings to the Renaissance master Leonardo da Vinci, for example, more than 4000 of his drawings survive [16].

The backgrounds of drawings not only vary considerably, given the highly differentiated qualities and properties of paper, but often dominate the visual field. Whereas the substrate surfaces of most paintings lie hidden beneath the medium, the foreground marks of a drawing may occupy a minuscule small portion of the overall area. As a result, even small differences in paper tonality may bias a classifier. Noting this hazard, Elgammal et al. [14] used both hand-crafted and learned-representation features to extract and represent each stroke in a drawing using neural networks. A key limitation of this approach is its inapplicability to works, such as those in chalk, that do not consist of discrete, extractable strokes.

Drawings are also the frequent subject of forgery. Beginning perhaps with Mauritius van Dantzig in the mid-twentieth century [17, 18], art connoisseurs have employed heuristics to distinguish forged from genuine artist marks. A hoary example is van Dantzig's contention that the spontaneous, fluid line of the “craftsman” can always be distinguished from the inhibited, excessively meticulous line of an imitator. Eric Hebborn, whose forgeries have fooled many experts and may number in the thousands, claimed to draw while slightly drunk to avoid this telltale propensity [19]. Any fixed principle for discerning the true stroke of an artist can likely be learned and spoofed by dedicated forgers, if not the artist's own students or admirers.

While the work we describe below points to the existence of a detectable pattern signature across an artist’s drawing oeuvre, it is not something that can be articulated in words or captured in a hand-crafted feature. Rather, our approach here, as in our previously cited work with paintings, was to trust in the pattern-recognizing capabilities of a convolutional neural network (CNN) trained on a corpus of images selected to best exploit those capabilities. Our contribution to existing work stems from the unique challenges posed by drawings — the varying, often dominating backgrounds, the flaws, and the prevalence of mixed media — and how we handled them. As detailed below, we found that classification accuracy strongly depends on preprocessing strategies and identification of an optimal subimage areal size for analysis. Surprisingly, we found it essential to train the CNN over a broad range of an artist’s drawing oeuvre rather than differentiating among tools and techniques, which might have been expected to yield greater accuracy within such categories. Most importantly, we found that a cross-validation methodology could be used obtain a confidence level when classifying an image: uniformity among predictions generated by identical models trained with overlapping but distinct image sets correlates with the accuracy of those predictions when averaged. This approach is essentially the reverse of ensemble techniques used as defenses to adversarial attacks. Whereas those systems combine predictions to defeat perturbations affecting one or a few of them, we exploit differences among predictions to assess their overall reliability.

2. Methods

2.1. Artist and Dataset

For this study, we chose as our subject the drawings of Raffaello Sanzio da Urbino (1483–1520), known as Raphael. Renowned as a painter and architect, he was also a prodigious draftsman, and a substantial body of his drawings survives. Raphael ran a workshop of assistants and students, and his work has been both imitated and forged — including by Eric Hebborn [19]. Raphael’s work reflects use of a wide range of drafting tools. This provided an opportunity to test whether segregating the work by technique would improve CNN performance.

Our training and test dataset consisted of 263 drawings, 104 of which are reliably attributed to Raphael [1]. Our 159 comparative images were chosen to span a range of similarity to Raphael’s work. They included works by close Raphael imitators (e.g., drawings labeled by their holding institutions as “Circle of [or similar qualifier] Raphael”), members of his workshop, and a known forger; drawings by Raphael's teacher, Pietro Perugino, and by earlier, contemporary, and later Italian artists, including Michelangelo and Leonardo da Vinci. We also included drawings by northern European and British artists, as well as by
nineteenth-century French admirers who drew "after Raphael."[2] Our objective in choosing these works was to train the CNN to make fine as well as coarse distinctions and to generalize beyond the training images.

We also compiled an additional "out-of-sample" set of 49 images: 27 by Raphael and 22 drawn by other artists. Here we deliberately chose many images having significant damage, irregular edges, staining, faded or faint artist marks, and other imperfections to test performance. We did not edit these out-of-sample images in any way — i.e., extraneous marks such as stamps and signatures were left undisturbed.

2.2 Image Preprocessing and Tiling

The drawings in our dataset varied widely in size — from a minimum dimension of 2.5 cm to a maximum dimension of 54.8 cm. We rescaled the images to a consistent resolution of 31 pixels/cm of the original drawing. We lightly edited training images to remove signatures, collector’s stamps, and large stains in order to avoid spuriously biasing the CNN analysis. All images were converted to grayscale.

To address the widely varying background tonalities, we first investigated simple contrast adjustments. We quickly learned, however, that the loss of detail resulting from increasing contrast impaired the CNN’s performance more than the varying backgrounds did. Moreover, the error resulting from different background tonalities averages out to some extent during training anyway. Both the Raphael and comparative images exhibited similar background variation, and the grayscale conversion eliminates background color as an error source.

To improve performance, we experimented with various forms of equalization. Commonly used to increase global image contrast, equalization raises the contrast where low but leaves higher-contrast regions unaffected. This results in less overall information loss than simple image-level contrast increases. “Histogram” equalization effectively partitions the intensity values of an image into histogram bins and spreads out the most frequent values. “Adaptive” histogram techniques redistribute lightness values based on several histograms, each computed for a distinct portion of the image. Contrast-limited adaptive histogram equalization (CLAHE) examines each histogram image zone and adjusts the contrast to avoid noise amplification and improve detail retention.

These equalization techniques operate on individual images. Because many drawing images consist mostly of background, we tried normalizing an entire set of images to a consistent background value. Specifically, we obtained whole-image histograms, averaged their peaks, and shifted the peak of each image to this average value. This approach succeeded in producing images with background regions that appear visually identical. Fig. 1 shows the effects of applying these various techniques to a representative Raphael drawing.

We processed the resized images into four separate grayscale sets: one without any modification, a background-normalized set, and sets produced using histogram equalization and CLAHE. Each image set contained all 263 drawing images. From each of these image sets, we prepared tiles using the procedure described in [11]. In particular, we decomposed the images into overlapping tiles ranging in size from 100x100 to 350x350 pixels in 50-pixel steps; larger tiles exceeded the shorter dimension of too many of the resized images. Tiles were sifted using the image entropy criterion given by:

$$H = - \sum_k p_k \log_2 (p_k)$$

where \(p_k\) is the probability associated with each possible data value \(k\). For a two-dimensional eight-bit grayscale image, \(k\) spans the 256 possible pixel values \([0..255]\). The 85-90% of tiles whose entropies fell below that of the source image were discarded. The image entropy criterion performed well in eliminating background regions, even those with significant motting and staining, while retaining the information-rich marks. As shown in Fig. 2, this was true for both chalk and pen images.

Tiles were overlapped sufficiently to produce, at each tile size, at least 12,500 total tiles from the Raphael images and a similar number from the comparative images. Given the high rejection rate, this required a considerable degree of overlap — 67% overlapping area for 100x100 tiles but 90% overlap for 350x350 tiles. The high degree of data redundancy does not preclude effective training as long as the image dataset is large and diverse enough.

To train our CNN models, we used four-fold cross-validation with tiles derived from each of the differently processed image sets serving as folds. For each image set, each of the four folds contained tiles corresponding to 198 training images and 65 test images with no redundancy among the test images in the different sets.

2.3 Deep Learning Model

We tested three CNN architectures in this study: the five-layer model described in [11], VGG16, and ResNet 50. As we found with analysis of paintings, a simple architecture is essential. Our five-layer model performed to expectations while VGG16 and ResNet 50, pre-trained on the ImageNet dataset or used without pre-training, performed poorly. The five-layer design includes five convolutional layers, five max pooling layers, five batch normalization layers, and three dropout layers. We employed sigmoid activation, a binary cross-entropy loss function, and an Adam optimizer. Source code for this model has been posted.[3]

We trained for 40 epochs in each cross-validation fold using a batch size of 16 and a learning rate of 0.0001. After each training epoch, the resulting model was saved. Tile-level prediction probabilities were averaged to produce an image classification, and accuracy in classifying the fold test images was assessed for each saved model. Averaging across tiles renders tile-level validation largely irrelevant to overall classification accuracy, so early stopping is not a useful strategy, nor are validation metrics particularly meaningful. Instead, optimal practice is to train for a sufficient number of epochs to obtain a distinct accuracy peak.
To determine scholarly consensus, we first consulted [1] and the information on the website of the drawing’s holding institution; we used the catalog of the 2017 Raphael drawings exhibition at the Ashmolean Museum [2] to review the most recent research on the works included in that show. As for all of our projects, we were able to obtain many large digital images from the websites of museums and other institutional repositories, and we are grateful to those organizations that make such resources freely available. We would like to extend special thanks to the Ashmolean Museum, repository of one of the largest collections of drawings by Raphael and his circle, for its generosity in sharing images.

Our complete datasets are available at https://sites.google.com/view/art-eye-d/datasets.

[3] [citation omitted to preserve author anonymity].

3. Results

3.1. Effects of Equalization and Tile Size

We explored variations in equalization technique and tile size in tandem. Training and testing through the four cross-validation folds at each tile size and for each approach to equalization (including no modification to grayscale images) revealed distinct advantages for 200×200-pixel tiles and CLAHE preprocessing. Perhaps not surprisingly given the muddy, noisy images produced by histogram equalization as exemplified in Fig. 1, this technique produced few usable tiles following entropy sifting and was not explored further.

Tiles of all sizes derived from the CLAHE images performed well across the in-sample folds. Although Fig. 3 suggests only a marginal advantage for 200×200-pixel tiles, in fact the superiority was more pronounced: in each of the four folds, perfect classification accuracy was attained after just a few epochs and largely persisted thereafter. At all other tile sizes, by contrast, peak accuracy was less than perfect or, in the case of 350×350-pixel tiles, occurred later and with fewer saved models.

Model performance was far different on the out-of-sample set, with lower accuracies and sharper differences among tile sizes. The accuracies illustrated in Fig. 3 reflect averaging across the best models from the four cross-validation folds; that is, each image was classified by the best-performing model from each fold, and the four resulting probabilities were averaged to produce a final classification for the image.

Although tiles produced from pure grayscale images or those preprocessed using background normalization resulted in lower accuracies relative to the CLAHE set, the 200×200-pixel tiles invariably elicited the best results relative to the other tile sizes tested. Why a particular tile size proves optimal for a given artist may be debatable; that one will emerge, however, seems inevitable, at least in our experience. This is true across artists and media, as well as for subject matter far afield of art [20]. For Raphael, it seems clear that limiting CNN attention to stroke-level detail cannot resolve distinctions between his drawings and those of others. Raphael’s distinctive “signature” as recognized by the CNN emerges at a larger feature level (see Fig. 2).

Holding tile size constant at 200×200 pixels, we can examine the effects of preprocessing technique. Fig. 4 shows results for histogram peak-shifted, CLAHE and pure grayscale tiles across the four folds for both in-sample and out-of-sample datasets. Table 1 reports the fold accuracy scores and standard deviations. CLAHE-preprocessed images produced not only higher maxima and mean values but also tighter spreads across folds. The latter feature suggests less accuracy-compromising noise, which may largely account for the superior performance.

<table>
<thead>
<tr>
<th>Equalization Type</th>
<th>Mean Accuracy</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histogram shift</td>
<td>0.68</td>
<td>0.067</td>
</tr>
<tr>
<td>CLAHE</td>
<td>0.77</td>
<td>0.019</td>
</tr>
<tr>
<td>None</td>
<td>0.57</td>
<td>0.106</td>
</tr>
<tr>
<td>Out-of-Sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histogram shift</td>
<td>0.87</td>
<td>0.03</td>
</tr>
<tr>
<td>CLAHE</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>0.77</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 1 – Mean accuracy scores and standard deviations across four cross-validation folds for in-sample and out-of-sample datasets using different forms of preprocessing

3.2. Chalk vs. Pen: One Class or Two?
We next investigated whether accuracy could be improved by segregating chalk (including charcoal) drawings from those made with a sharp tool, which we collectively refer to as “pen” drawings. The possibility seemed realistic given the clear differences in mark appearance and application technique. Indeed, our initial assumption was that the visual differences might be so pronounced as to require separate classes. As we assembled the dataset, however, we found that many works combine both types of media in varying proportions. Consequently, we knew that efforts to identify a dominant medium would involve subjective judgment and risk biasing the results.

Using all of the pen-dominant drawings from the curated dataset and supplementing these with additional pen-dominant drawings produced a set of 169 images (74 Raphael and 95 comparative), while for chalk-dominant drawings the set contained 94 images (30 Raphael and 64 comparative). These datasets were considerably smaller than our curated set. To preserve a sufficient number of test images, we used three-fold (rather than four-fold) cross-validation for the chalk-dominant set.

After CLAHE preprocessing, we overlapped tiles sufficiently to achieve at least 12,500 total Raphael tiles and a similar number of comparative tiles. The smaller numbers of source images required greater degrees of overlap than were necessary, at 200×200 pixels, for the curated dataset. Whereas the latter required 80% overlap, the overlaps necessary for the pen and chalk datasets were closer to 90%.

We need not have worried about bias in identifying a dominant medium. The results obtained with separate chalk and pen datasets were inferior as set forth in Table 2.

<table>
<thead>
<tr>
<th>Drawing Set</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In-Sample</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pen-dominant</td>
<td>.84</td>
<td>.03</td>
</tr>
<tr>
<td>Chalk-dominant</td>
<td>.80</td>
<td>.07</td>
</tr>
<tr>
<td><strong>Out-of-Sample</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pen-dominant</td>
<td>.61</td>
<td>.07</td>
</tr>
<tr>
<td>Chalk-dominant</td>
<td>.47</td>
<td>.02</td>
</tr>
</tbody>
</table>

Table 2 – Mean accuracy scores and standard deviations for pen-dominant and chalk-dominant datasets

The models trained solely on chalk-dominant or pen-dominant images sharply underperformed models trained on the mixed curated image set. By comparison, as reported in Table 1, averaging across the best fold models trained on the curated image set produced an out-of-sample classification accuracy of 0.77. In particular, of 48 tested images in the mixed-medium out-of-sample set, 37 were classified correctly (averaging across the best fold models trained on the curated image set) and 11 incorrectly. The 11 incorrect classifications were evenly split — six chalk-dominant, five pen-dominant — and all were false positives. The mixed out-of-sample set was itself roughly split between pen-dominant (25) and chalk-dominant (23) images.

The models trained on chalk-dominant or pen-dominant images were tested only against pen-dominant or chalk-dominant out-of-sample images. Why do models trained on a dataset that includes both pen-dominant and chalk-dominant drawings perform so much better than models trained on datasets that separate the two? Indeed, for models trained on chalk-dominant drawings, the results on the out-of-sample set were worse than guessing.

Certainly, part of the answer lies in the smaller size of the chalk-dominant dataset. But the large performance disparity suggests other, more influential factors at work. Chalk drawings, particularly those that are centuries old, have far more variability in quality of mark than pen drawings due to the greater material vulnerability of chalk. The mark character of chalk also varies more, from soft and broad to bolder and finer strokes. The Viti drawing illustrated in Fig. 2, for example, is diffuse throughout with broad marks. It was misclassified by all CNNs tested, including those trained on the mixed dataset. Such visual features may simply have too little differentiation among artists drawing in a similar style to be resolved by the CNN.

Moreover, as noted, many drawings contain both chalk and pen passages. An artist’s signature style may reflect not only how specific media are applied but how they are combined. If training images are skewed toward chalk-like or pen-like drawings, the CNN will only rarely encounter the combined media during training. This effectively removes a signature characteristic — i.e., a degree of freedom — from consideration relative to the more broadly trained CNN.

c. Behavior Across Cross-Validation Folds

It is well known that CNN performance can be affected unpredictably by various forms of image noise, even in small amounts [21]. Indeed, this unpredictability is often exploited to undermine face recognition and other CNN-based detection systems [22]. CNN-spoofing noise sources can include lighting variations, blur, differences in camera sensors, and contrast variations [21,23]. We found that we can exploit the unpredictable sensitivity of CNNs to atypical conditions to reveal the accuracy-compromising existence of those conditions. We can then better understand the inherent accuracy of the prediction.

Because each of the best cross-validation models was trained on a different but largely overlapping mix of training images, their predictions should be similar for a new image. The number of training images in each fold seemed sufficient, given our experience with other artists, to support good (and consistent) model performance. We verified this by training new models on the entire in-sample image corpus and testing against the out-of-sample set. Although this
carries some risk of overfitting to the out-of-sample set, the results reflected only a modest improvement (from an accuracy of 0.77 to 0.83) over averaging the best cross-validation models. Those models, in other words, performed well enough individually that their predictions should be similar.

Various factors may account for divergence among model predictions for a given image. These may arise from characteristics of the drawing itself (due to rough handling and the passage of time, as described above), or can stem from the training underlying the CNN models. In the former case, the image may contain noise or blur not immediately apparent on visual inspection, and because of the unpredictable CNN sensitivity to such defects, different models respond differently. In the latter case, the training set may simply lack enough differently labeled images with sufficient visual similarity to the work under study to classify it reliably. Either way, the degree of divergence—the spread between prediction extrema—strongly correlates with the accuracy of the averaged prediction. For example, a Hebborn forgery in the out-of-sample set was weakly classified (with an averaged probability of 0.54) as drawn by Raphael. But in fact, the classification probabilities ranged from 0.18 to 0.75, a very wide range clearly indicating failure. Had the training set included more Hebborn (or Hebborn-like) images, the output probabilities may have converged toward a correct classification.

<table>
<thead>
<tr>
<th>Spread</th>
<th>Surviving Image Fraction</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 0.25</td>
<td>.75</td>
<td>.81</td>
</tr>
<tr>
<td>≤ 0.20</td>
<td>.63</td>
<td>.83</td>
</tr>
<tr>
<td>≤ 0.15</td>
<td>.5</td>
<td>.92</td>
</tr>
<tr>
<td>&lt; 0.14</td>
<td>.44</td>
<td>1.0</td>
</tr>
<tr>
<td>≤ 0.10</td>
<td>.33</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3 – Effect of spread between minimum and maximum probabilities across best fold models on classification accuracy. The surviving image fraction represents the proportion of images whose spread satisfies the constraint. Accuracy reaches 100% when the spread is below 0.14.

As indicated in Table 3, when the probability spread falls below 0.14, all qualifying images are classified properly. The price of increasing accuracy, however, is exclusion of images we might wish to study—in the case of our out-of-sample set, more than half of them in order to achieve 100% accuracy. Of course, the fact that the out-of-sample set includes many noticeably flawed images makes it something of a worst case.

Even with the largest probability spreads, classification accuracy never falls below 0.77. Table 3, although based on a limited number of out-of-sample images, suggests how inherent accuracy varies with the spread. The averaged classification probability assigned to a candidate image can be qualified by the accuracy limit corresponding to the observed spread to produce a more meaningful, adjusted prediction. In addition, we found all classification errors to be false positives. As a result, these inherent accuracy limits do not directly apply to classifications of candidate works as not by Raphael, since our models have yet to produce a false negative.

4. Conclusion

We have demonstrated the feasibility of classifying drawings computationally in order to judge attribution and authenticity. Many complicating factors, ranging from spurious marks to widely varying paper characteristics to the numerous forms of cumulative damage, can be addressed with suitable image preprocessing and tile sifting. Beyond these automated operations and selection of images, we exert minimal influence over the CNN and allow it to guide our efforts rather than vice versa. It is the CNN that determines the optimal tile size for analysis and which pattern features promote successful classification. The CNN performs best when presented with carefully curated works across an artist’s entire oeuvre, without arbitrary exclusions based, for example, on drawing tools or technique. Our cross-validation methodology provides a basis for assigning a confidence level. This inherent accuracy limit, specific to an image under study, would remain hidden were a single model employed.

Figure 5 illustrates a system-level view of the preprocessing and analysis steps we describe. All of the depicted steps are fully automated; there are no handcrafted features, nor is there any manual review of images or data. After the cross-validation models have been trained and selected, the steps of image
preprocessing, tiling, sifting, and analysis may be performed on sequences of candidate images in a workflow process. A collection of images under study, for example, may be uploaded to a server for rapid classification, reporting any qualifications or uncertainties. Indeed, the simplicity of the CNN architecture employed in this study allows for deployment directly on mobile devices. At 200×200 pixels its parameter count is 438,385, about 1/10 that of the MobileNet v2 architecture expressly designed for mobile use.

The increasing difficulty of obtaining expert opinion on questions of authenticity is well-documented; art experts fear loss of reputation as well as loss in the courtroom [24–26]. Our system offers the possibility of a convenient “first look” at attribution: one that may suggest the utility of professional expertise and perhaps even encourage otherwise reticent experts to venture their opinions.

References


18. M.M. van Dantzig, Vincent?: a new method of identifying the artist and his work and of unmasking the forger and his products, Keesing, Amsterdam, 1953.


Competing interests:

The authors declare no competing interests.

Figures

Figure 1

Raphael, *Head of St. James* (c. 1498-1520). Clockwise: original image; grayscale with no adjustment; histogram-equalized image; CLAHE image; peak-shifted image. Because the background of the original image was of average tonality, the peak-shifted image differs little from the grayscale image with no adjustment.
Figure 2

Top, Raphael, *Sketches of Cattle* (c. 1512) (Pen and iron-gall ink) and 200×200 tiles derived from grayscale version; bottom, Timoteo Viti, *The Virgin and Child (after Raphael)* (c. 1484-1523) (black chalk) and 200×200 tiles derived from grayscale version. For both drawings, the illustrated tiles are the only ones that survived sifting with the image entropy criterion. The tiles capture the visually busy areas of the drawing rather than the background.

Figure 3

Performance comparison among tile sizes

with tiles generated from CLAHE-preprocessed images
Figure 4

Box-and-whisker plots showing extremes, quartiles, and median values across folds for 200×200-pixel tiles of in-sample (top) and out-of-sample (bottom) images. In both cases, CLAHE preprocessing results in better accuracy and a narrower spread.
System-level view. A source image, suitably downscaled to the resolution used for training, is converted to grayscale and preprocessed with CLAHE. Overlapping tiles are generated from the source image and sifted, with a sufficient number surviving to support classification (typically no more than dozens are required). Classification probabilities associated with qualifying tiles are obtained and averaged using multiple cross-validation models trained on different but overlapping image sets. If the averaged tile-level classification probabilities for the cross-validation models fall within an empirically determined maximum spread, these average probabilities are themselves averaged to produce a final prediction probability.