Large scale style based dating of medieval manuscripts

Fredrik Wahlberg
Dept. of Information Technology
Uppsala University
fredrik.wahlberg@it.uu.se

Lasse Mårtensson
Dept. of Business and Economics Studies
University of Gävle

Anders Brun
Dept. of Information Technology
Uppsala University

ABSTRACT

In this paper we propose a novel approach for manuscript dating based on shape statistics. Our goal was to develop a strategy well suited for a large scale dating effort where heterogeneous collections of thousands of manuscripts could be automatically processed. The proposed method takes the gray scale image as input, then uses the stroke width transform and a statistical model of the gradient image to find ink boundaries. Finally, a distribution over common shapes, quantified using shape context descriptors, is produced for each manuscript image. The proposed method is binarization-free, rotational invariant and requires minimal segmentation.

We evaluate our work on the 10000+ manuscripts collection "Svenskt diplomatariums huvudkartotek", consisting of charters from the medieval period of todays Sweden. The images, originally intended for web viewing, were of low quality and had compression artifacts. Due to unsupervised feature learning and regression, the collection could be dated with a median absolute error below 19 years even though we only used 5% of the labels in the estimator training.

1. INTRODUCTION

When estimating production dates for manuscripts in the large collections from the late medieval period still in existence, the most common and trusted way is using linguistics and palaeography. Lately automatic writer identification techniques have been applied to dating for improving the scalability of the dating process.

We propose a feature extraction scheme based on the stroke width transform, the shape context descriptor and bag-of-features matching. The stroke width transform was used as a denoising strategy, to make sure edge pixels always belonged to pen strokes. These edges were then analysed using the shape context descriptor, creating a codebook of common shapes. A bag-of-features approach was then used to create a distribution over shapes on a manuscript page. The result

Figure 1: Four charters from the collection "Svenskt Diplomatariums huvudkartotek" (SDHK), on which all dating in this paper was performed (section 1.1). Note the different kinds of degradations in the material (e.g. staining, holes and varying contrast). The resolution was "web" quality, in this case 1.5 **Mpix for each image.**

was a feature space where the symmetric Kullbach-Liebler distance was interpreted as a pair-wise dissimilarity metric between manuscript pages. Employing Gaussian processes and support vector machines in this feature space, we show that as little as 5% of the data needs to be labeled for a dating on par with a human expert.

A great problem facing researchers in automatic writer identification, dating and text recognition is the low quality images. Often, librarians and historians have different intentions for the images than computer scientists when digitizing the source material. Having readable text and a low image size might be optimal for a library, but not for image analysis of the same material. We have developed our methods for the low resolution images of the database SDHK, presented below, and show that high image quality is actually not needed using our method.

When researching writer styles (e.g. for identification) in a forensic research scenario, most or all of the data is labeled already by law enforcement personnel. The research problem might be to associate a piece of handwriting to a set of known writers (i.e. finding the culprit). In historical applications, the labeled material is scarce and the scholars able

to do the labelling sometimes even more so. Also, the signing author was often not the scribe (maybe the most famous example is the scribe Marcus Tullius Tiro, a person owned and freed by Cicero, and likely the inventor of the Tironian nota). We wanted to use as little training data as possible and have here set the limit to using 5% of the labelled data for training.

An exciting outcome of our work is that we have found errors in the ground truth when studying it. Charter 258 was initially an outlier in our dating until we discovered that it was actually correctly dated by the proposed system but mislabeled in the ground truth by 200 years. Another outlier turned out to be a photograph of a $17th$ century copy of the medieval charter.

1.1 Svenskt diplomatariums huvudkartotek

The collection "Svenskt diplomatariums huvudkartotek" (SDHK) is a collection of charters from the medieval period in the geographic region of present-day Sweden. The collection have been photographed during a period of 10 years and is still ongoing. The number of low quality images we have obtained is over 11000. All of the images and meta-data we used are available without charge on the website of the Swedish national archive[1] with a low resolution (1.5 Mpix) and jpeg compression ($\approx 85\%$). We have chosen to use 11000 of these images that we deem to have a reasonable image quality. Some images from the collection are shown in 1. The images we did not use were from an older reproduction project which made them harder to binarize, something we need for the comparison with earlier work.

The medieval era in the geographical region spans from the end of the viking era (1050) until the coronation of King Gustav I (1523). A very nice property of these charters are that most are dated on the day. This makes SDHK the perfect database for evaluating computer assisted dating algorithms since we have approximately 11000 samples from multiple writers from a period of almost 500 years with perfect ground truth (an accuracy far less than a year).

The charters comes from all over medieval Sweden but also from other places in Europe (e.g. Rome and Paris). It is unlikely that a small set of scribes wrote a significant portion of the collection and hence, the risk of but accident classifying single hands instead of style based dating is small. However, it is likely that many of the medieval scribes whom we have manuscripts from outside of SDHK are also represented in SDHK and hence it could be used for writer identification in the future. What gives the endeavour described here such importance, is the possibility to date manuscript where the palaeographical, historical and linguistic features have failed to give a reasonable estimate of the date of production.

1.2 Previous work

In [7], a dating approach based in the Hinge ([13]) feature was proposed. The Hinge feature has been shown to work in writer identification before and gave satisfactory results also in manuscript dating. The authors also used a two step support vector regression approach to finding the date. The method is evaluated on a collection of Dutch charters from the years $1295-1555$, collected in 11 bins with charters form the years $\{1300\pm5, 1325\pm5, 1350\pm5, \cdots, 1525\pm5, 1550\pm5\}$. Below, we will present results using the Hinge feature as a comparison to our proposed method, but without binning the data.

The shape context descriptor, proposed in [1], have been used on handwritten material and related tasks in the literature. In the original paper, the descriptor is used to find a deformation grid aligning two instances of an object. The authors show the relevance of shape context for a number of interesting applications (among them the famous MNIST database with handwritten numbers). The shape context descriptor is robust to small variations in the shape (e.g. scaling) and can be made rotational invariant.

In [9], an extension of [1] was proposed where a distribution over a codebook with shape context descriptors was used as a feature. This was later used in [12], where the authors classified images of company logos. They extracted shape context descriptors from the edges of the logos, create a codebook and compared the distributions over descriptors for the logo images. To reduce the search time the authors employed a locality sensitive hashing strategy, approximately solving the nearest neighbour problem while still keeping the computational demands low.

2. METHOD

Below, we propose a novel approach for dating premodern manuscript. Through the use of shape context vectors extracted from the edge map of the manuscript images, we created distributions over a number of common shapes for each page. A codebook was learned in an unsupervised manner from the image collection, determining the base shapes for the aforementioned distribution over common shapes.

Our intention is to publish our source code when fully evaluated. The current implementation stands on the shoulders of Python SciKit Learn [10] and GPy [16].

2.1 Feature extraction pipeline

We describe the proposed feature extraction pipeline below. A flowchart of the process is shown in figure 2.

2.1.1 Edge distribution model

When performing a segmentation, a significant amount of domain knowledge often goes into the algorithm development. If assumptions made do not hold, the segmentation often fails. The feature development presented here was intended to solve a dating task on a very heterogeneous collection of manuscripts (as shown in figure 1). We have identified edges of ink strokes using a statistical model for the edge magnitudes and the stroke width transform.

The stroke width transform takes an edge image as input. This is commonly supplied by the Canny edge detection algorithm ([2]). However, Canny needs two threshold parameters for performing the edge detection. We used a statistical model of the distribution over gradient magnitudes in the image to automatically find the two thresholds, rendering our approach parameter free in this step.

[1] riksarkivet.se/sdhk

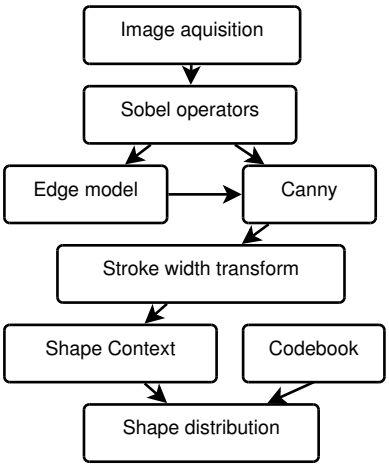

Figure 2: A flowchart description of the feature extraction process. The different stages of the pipeline are described in detail in section 2.1.

To estimate the two parameters needed for Canny edge detection, we modeled the distribution over gradient magnitudes, created using Sobel operators and the l_2 norm, with a Gaussian mixture model (GMM). The GMM was set to have two components, one for the background "spike" of very low gradients and one for the stronger magnitudes. The GMM was defined as in equation 1 (where x is a gradient magnitude).

After fitting the GMM to the magnitude data, the thresholds could be estimated. The lower threshold (T_{low}) was defined as the point at which the likelihoods of the two mixtures were equal (shown in equation 2). The higher threshold (T_{high}) was defined as the higher mean value of the two components (as shown in equation 3).

$$p(x) = \omega_1 \mathcal{N}(x \mid \mu_1, \sigma_1^2) + \omega_2 \mathcal{N}(x \mid \mu_2, \sigma_2^2) \quad (1)$$

$$\omega_1 \mathcal{N}(T_{low} \mid \mu_1, \sigma_1^2) = \omega_2 \mathcal{N}(T_{low} \mid \mu_2, \sigma_2^2) \quad (2)$$

$$T_{high} = max\,(\mu_1, \mu_2) \quad (3)$$

The gradient magnitude distribution and the final estimation of the thresholds are illustrated in figure 3.

2.1.2 *Stroke width transform*

The stroke width transform (SWT), proposed in [4], is a transform where each pixel in an image gets labeled with the most likely stroke width at that point. The SWT was initially applied for spotting text in image sequences where patterns for likely width were learned from a database. The stroke width is estimated by following a line in the gradient direction from each edge pixel (in to the potential pen stroke area) and finding the euclidean distance to the closest edge pixel with an approximately opposite gradient ($\pm 45°$). An example is shown in figure 5.

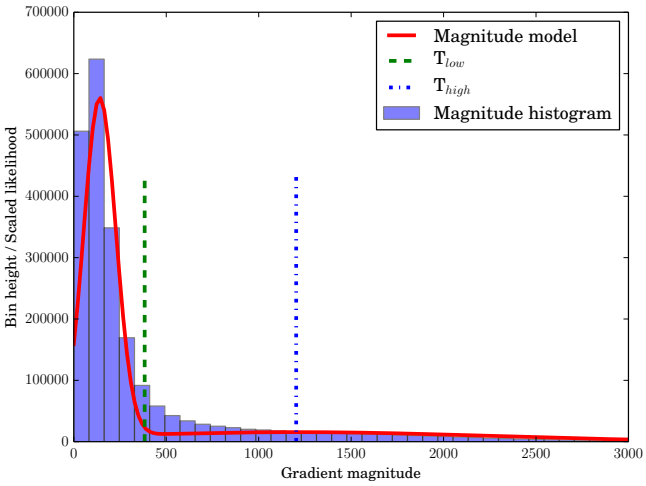

Figure 3: A histogram over gradient magnitudes of a sample page. The red line indicates the likelihood (scaled for comparison) of the edge magnitude distribution model from section 2.1.1. Note the estimated thresholds for the edge detection, shown by the dotted vertical lines.

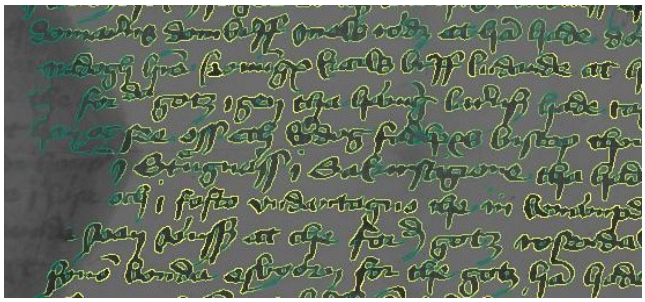

Figure 4: A grey scale text image from SDHK with the edges (weighed by magnitude) superimposed in colour. The edges were extracted using Canny edge detection with the parameters automatically identified using a Gaussian Mixture Model based approach. The more yellow edges are those with a stronger gradient magnitude and the more green ones those with a weaker gradient magnitude.

We used the SWT for denoising with the assumption that all edges with no opposite edge did not belong to ink strokes, assuming the edge detection worked properly. In figure 4, the result of this process is show on a manuscript page.

2.1.3 Shape context

The shape context descriptor was proposed in [1]. The descriptor was extracted from a binary image patch (of a set size) and described the distribution of the foreground pixels. The coordinates of the foreground pixels in polar coordinates (centred around the middlemost pixel in the patch) were collected into a histogram. The radius was described in log space, giving more importance to close pixels than those further away. The dimensionality of the descriptor was the number of bins for the angle (8) times the number of bins for the radius (5).

In a patch P, each foreground pixel coordinate pair $(x_i, y_i) \in P$ was transformed as $\theta_i = \text{atan2}(y_i, x_i)$ and $r_i = \log \sqrt{x_i^2 + y_i^2}$ into the new coordinate pair $(\theta_i, r_i) \in P$. The polar range of the histogram was normalized to the span of the values possible for any log polar coordinate in the patch. The number of angle bins q and the number of log radius bins p are model parameters. We have preliminary results indicating that varying the p, q and N (patch side) give very similar performance, though this has not been thoroughly tested.

Another way of describing the same process is illustrated in figure 6. In a large collection it can not be assumed that all writing is horizontal in the image. We achieved rotational invariance by subtracting the gradient direction of the centre pixel from all coordinates in each patch.

2.1.4 Codebook training

In [5], the authors proposed matching images using a collection of unordered features, a so-called bag-of-features. We find creating a bag of unordered features is a very intuitive approach to dating. Statistics on which forms were common over time is used as a feature in palaeography. However, in paleography they are used as a part of an analysis of the morphology of specific characters. Also, if the codebook is large enough, codebook entries signifying noise (or irrelevant features) could potentially be found.

In [9], the concept of shapemes was introduced. The shape context descriptors extracted from a figure was quantized using a codebook over common shape context descriptors. In this way a histogram over the descriptors in the codebook could be constructed to get a low dimensional representation of the distribution of descriptors in an image. We use a very similar approach to the shapemes of [9]. A codebook was created by sampling the training set of shape context descriptors from a set of images. The set of training images were chosen so the distribution of dates would be as even as possible. Note that creating a codebook is unsupervised learning and hence, more data than the human labeled data can be used.

In [14], a k-means algorithm suitable for large scale implementations was proposed. We have used this mini batch k-means (choosing a division of 20 bins for our training set) to get a codebook for the training set of descriptors. This algorithm created an approximation of a k-means codebook

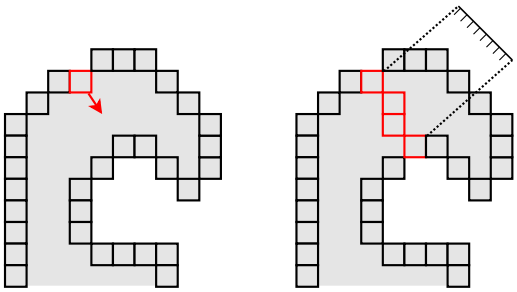

Figure 5: Illustration of the stroke width transform that was used for denoising in the proposed feature extraction. In the figure, cubes illustrate edge pixels and the shaded parts the inner areas of the ink strokes. The red cube to the left is the element being processed with the gradient direction as the red arrow is pointing. The right part of the figure is the final result. There a Bresenham line has been found, ending at an edge pixel with an approximate opposite gradient direction. The euclidean distance between the pixel centers is measured (illustrated by the small ruler), giving the final estimate of the stroke width at the edge pixel being processed. This procedure was run for every edge pixel. All edges with no opposite edge were removed.

(very accurate approximation, as show in the original paper) but was significantly faster.

In the end we got a codebook with a set dimensionality to compare each manuscript page against. The final feature was a distribution over the rotational invariant and scale insensitive shape context descriptors in the codebook.

2.2 Regression

For evaluating the pure effect of the feature, an unweighted k-NN regression was used. Using k-NN is a very naivë approach and was chosen because of this. A simpler approach leaves more of the "work" to the feature. To strengthen our argument, many competitions on writer identification also use k-NN for feature evaluation (e.g. [8]).

Also, to minimize the required work of a human expert, the training set was chosen to be small. When working with historical manuscripts, we seldom have the luxury of large quantities of labeled data. Hence, k-NN is unsuited for our purposes (since it usually requires exactly this). Below we will also evaluate how well our proposed feature approach works with more advanced types of regression.

2.2.1 Dissimilarity metrics

All herein described features represent the document characteristics as a joint probability density over the different feature components. Though in some cases the feature was sampled from a continuous density, the resulting feature was a discretization of said density. In [17], we evaluated three distance metrics for discrete densities from a survey of dissimilarity functions for probability densities ([3]) . We concluded that χ^2 (shown in equation 4) and Jeffreys (shown in equation 5) were the most useful.

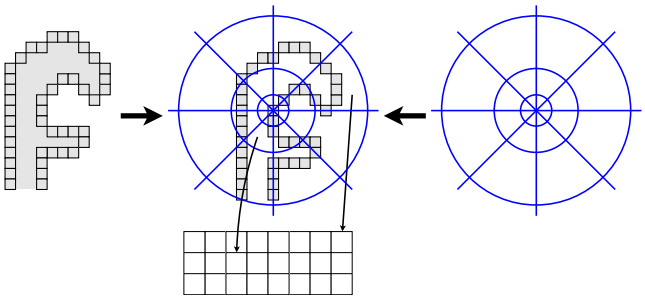

Figure 6: Illustration of the scape context descriptor. The upper left figure is a part of some letter with the contour pixels illustrated as box and the interior in gray. The upper right stencil represents the bins of the shape context descriptor in (note the round shape because of the polar coordinate conversion). The aforementioned figures a re overlayed and the edge pixels mapped to bins, creating a histogram below the overlay.

A problem with the χ^2 distance in equation 4 was that if $\mathbf{v}_i + \mathbf{u}_i = 0$ for any valid i in the summation, the division and, by extension, the full sum was undefined. Hence, we chose to use the Jeffreys divergence as the only distance metric for this study since it does not suffer from the above mentioned drawback.

$$D_{\chi^2}(\mathbf{v}, \mathbf{u}) = \sum_{i=0}^{|v|} \frac{(\mathbf{v}_i - \mathbf{u}_i)^2}{\mathbf{v}_i + \mathbf{u}_i} \tag{4}$$

In equation 5 we show the Jeffereys divergence between two discrete probability densities $\mathbf{v}$ and $\mathbf{u}$. It is based on adding up two Kullback-Leibler divergence functions to get a symmetric divergence.

$$D_{Jeffreys}(\mathbf{v}, \mathbf{u}) = \sum_{i=0}^{|v|} (\mathbf{v}_i - \mathbf{u}_i) \ln \frac{\mathbf{v}_i}{\mathbf{u}_i} \tag{5}$$

2.2.2 *Gaussian process regression*

To improve on the results from the k-NN approach to finding the dates we used Gaussian process (GP) regression ([11]). A Gaussian process is a stochastic process that we used for modelling the evolution of the writing style distribution over time. For every point on the timeline, a Gaussian distribution in the feature space was trained using the features from the training set images. In this model, the training feature vectors are seen as drawn from said Gaussian process. The regression part of this model is that the drawn features belong to a Gaussian distribution whos shape is dependent on time. A set of dated training feature vectors were used to create a model that maximized the likelihood of generating the training set.

A Gaussian process lets us find a distribution in the feature space where we by taking a "slice" of the distribution for any date can get an estimated distribution of features at that date. We used this for finding the position in time where a manuscript page would fit with the maximum likelihood. When the independent date variable was set, the result was a Gaussian distribution with the same dimensionality as the feature space. In a sense, a GP is an infinite dimensional Gaussian distribution and is defined at every value the independent variable can take.

The covariance of a GP is defined by a kernel function. If this kernel have parameters, they are estimated from the data to maximize the likelihood of the GP generating the training data. As the kernel function we chose a radial basis function (RBF) with automatic relevance determination (ARD) shown in equation 6. Using ARD gave us the possibility to "stretch" the feature space.

$$K_{ARD}(u, v) = \exp\left(-\sum_{i=0}^{|u|} \gamma_i (u_i - v_i)^2\right) \tag{6}$$

For determining the ARD weights (γ in equation6), the gradient of the log likelihood was estimated and optimized using gradient methods. Using an ARD kernel made it possible to re-weight which shape context descriptors (in the codebook) would be more important for estimating the date (and even remove shapes that were irrelevant for the task). The implementation by [16] was used as a base for our approach.

2.2.3 *Multi step support vector regression*

In [6, 18], a multi step support vector regression (SVR) approach was developed that included local re-estimations of the regression to increase accuracy. This approach was also used in [7] for dating medieval manuscripts.

At first a global SVR was trained on the training set of images to create a model for the development of the writing style over time. The chosen kernel was a standard radial basis function as shown in equation 7. The parameters were found using a grid search where $\gamma \in \left\{2^{-15}, 2^{-14}, \cdots 2^{14}, 2^{15}\right\}$ (RBF parameter), $C \in \left\{2^{-15}, 2^{-14}, \cdots 2^{14}, 2^{15}\right\}$ (regularisation parameter) and $\epsilon = 0.1$ (margin parameter).

$$K_{RBF}(u, v) = \exp(-\gamma \left\|u - v\right\|^2) \tag{7}$$

After the global model estimation, a local model was trained for each query manuscript (i.e. every document from the test set) to increase the accuracy. To create the local model, a query manuscript was dated using the global model. The local training set was defined as the intersection between the set with the 100 closest (in time) manuscripts from the query manuscript and all manuscripts (from the training set) within a time span of ± 2 of the mean absolute error of the global model (evaluated only on the test set).

To improve the result even more, we added a small extra grid search between the global and the local one to increase the accuracy of the global model. The final result was an estimation of the year in which the manuscript was produced.

3. EVALUATION

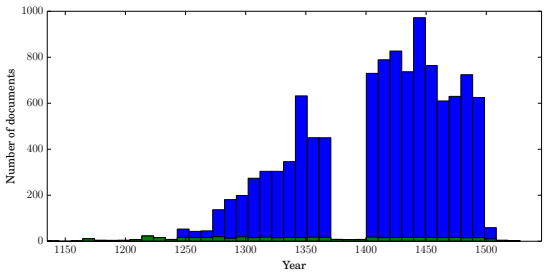

Figure 7: Histogram over the distribution of dates in the SDHK database (section 3.1). The blue bars is the distribution for the full database and the smaller green bars the distribution of the training set. Note that the training set is sampled as evenly as possible over time. Fewer charters have survived from the earlier middle ages and it is also likely that fewer were produced then. There is a gap in the late $14th$ century. This gap is due to low quality in the earliest photographs produced by the archive.

3.1 DB: "Svenskt diplomatariums huvudkartotek" (SDHK)

Our evaluation database was the "Svenskt diplomatariums huvudkartotek" (SDHK) described in section 1.1. In figure 7, a histogram showing the distribution over dates is shown. The collection spans from the $12th$ century to the $16th$ but most of the material was from the later years. The total number of binarizable charters was 10994.

The full set of manuscript images was split into a training set consisting of 500 manuscript evenly sampled over the time line and a test set with the rest of the images (10494). The training set was sampled evenly over the timeline so as to let the regression models catch the trends over time. The collection is dominated by a large numbers of manuscript from $1350-1450$ so a random sample would probably mostly get data from that smaller period.

3.2 Evaluation metric

As the main error metric we used mean squared error (MSE). It is shown in equation 8 where X is the estimated dates and X^* the real dates form the ground truth.

$$MSE = \sum_{i=0}^{|x|}(X_i - X_i^*)^2 \tag{8}$$

A desirable property in this metric is the extra emphasis on outliers. In our application an accurate estimate is important, but not as important as reducing the large errors. For the historians, an error of ± 30 is more than acceptable (in most cases). On the other hand, delivering a system to a humanist research group that have more than occasional large errors run the risk of severely damaging the trust in the system. This also ties in to the importance of giving an estimate of the uncertainty that we make possible by using a Gaussian process for estimations.

	P25	P50	P75	MSE	RMSE
Proposed 1-NN	13.0	32.0	62.3	4721	68.7
Proposed 5-NN	13.8	27.9	54.3	3145	56.0
Proposed GP	11.4	24.5	42.8	1719	41.5
Proposed SVR	**7.9**	**18.3**	**36.8**	**1389**	**37.3**
Hinge 1-NN	11.0	32.0	59.0	5172	71.9
Hinge 5-NN	10.6	26.8	51.5	3165	56.3
Hinge GP	15.9	33.6	60.2	5035	70.9
Hinge SVR	8.6	19.6	40.6	1998	44.7

Table 1: Best results, measured in MSE, for the different feature and regression types. The metrics presented are absolute error at the $25th$ (P25), $50th$ (P50) and $75th$ (P75) percentiles together with mean square error (MSE) and the square root of the former (RMSE). All numbers are measured in years except MSE. The numbers in bold font are the best results in their respective categories. The 1-NN and 5-NN are evaluated only on the 500 manuscript training set while GP and SVR are trained on the training set and evaluated on the 10494 manuscript test set.

Another convenient property is the MSE metrics relation the root mean square error metric (RMSE), which simply is the square root of MSE. This is closely related to the standard deviation of the error distribution (equivalent if it is Gaussian and there is no bias), giving an intuitive measure in years.

We also present $25th$, $50th$ and $75th$ percentiles from of the distribution of absolute estimation errors. We do this to give the reader more information on the distribution of absolute estimation errors for a particular feature evaluation set up.

3.3 Comparison

As a comparison for the proposed method, we have chosen the Hinge feature. It was originally presented in [13] and used recently for dating a set of charters from the Netherlands comparable to our collection (but much smaller) in [7].

The Hinge feature is dependent on a good binarization and to make a fair comparison we have tried to find a binariation method suited for our collection. After trying out different binarization methods, we use [15] instead of the original Otsu method used in [7].

The Hinge feature is a histogram over angles along the contours of the connected components. We have varied the parameters N_{leg} (used for estimating angles) and p (the side of the quadratic histogram). The histogram is symmetric around the diagonal and almost half of the bins are therefore redundant. We remove the redundant information leading to a dimensionality of $\frac{(p+1)p}{2}$ (in [7] the final dimensionality is 696). Also, we used bi-linear interpolation when assigning sample points to bins.

3.4 Results

In table 1, the best results for the proposed feature and the comparison are shown using 1-NN, 5-NN, GP and SVR. The proposed method works well despite the low quality of the

images. Classifying 75% of the charters within ±36.8 years is very good from a humanist standpoint. The P25 values look very good for all configurations, though it could be argued that errors below ±10 years does not matter. Such a low error should be hard to estimate on stylistic grounds since scribal hands likely did not evolve much in that time frame. The P50 and P75 is more important since the values are closer to the length of a scribes active years. Figures showing the error distribution are shown below.

In figure 8, some performance data relevant to the robustness of parameters are shown. The MSE is plotted against codebook size and also the 5 best parameter sets for the Hinge feature with SVR estimation. The proposed feature performs better for higher codebook sizes while still keeping the dimensionality well below the comparison. The 1-NN is very erratic, likely because of which shape context descriptors are a part of the codebook.

In figure 9, the moving error estimation is shown. At every point on the timeline, the average absolute error withing a timespan of ±20 years was estimated. In the k-NN, the error increases towards the ends of the time line. This is in proportion to where the main parts of the data material is. However, the evaluation for k-NN is only run on the evenly sampled training set implying that the variation in style might be larger in those areas. The 80% confidence is included to shown the variation, even though that metric assumes Gaussian data. In 10, the actual distribution of error is shown and it is very close to Gaussian.

In figure 10, the distribution of the errors are shown for k-NN and after regression. The distribution for the SVR has a clear bias while the GP is centred around 0, though the SVR performed better overall. The Gaussian and student-t curves are included to give some reference shapes in the figures.

4. CONCLUSIONS

We have introduced a novel feature extraction approach for dating medieval manuscripts. It was evaluated on the database SDHK, containing 11000 manuscript pages spanning from the 12th to the 16th century. All images were low resolution (1.5 Mpix) and with jpeg compression (≈ 85%).

The dating effort was started because of the need for digital methods in humanist research in Sweden. We have made a point of only using a 5% subset for training our estimators to show how our method could generalize the model to a large dataset. In the end, digital methods need to be applicable to the very real problems with large collections in historical research where experts are often few and far apart.

Some advantages to the proposed method are inherited from the shape context descriptor. Rotational invariance, robustness to small scale changes are the main ones. We have preliminary results indicating a robustness to the choice of shape context descriptor parameters. We have developed an edge model to make the method binarization-free. This also makes our method more robust to low quality images since we are not dependent on identifying foreground vs background. The dimensionality of the feature is chosen by selecting an appropriate size for the codebook. We show that

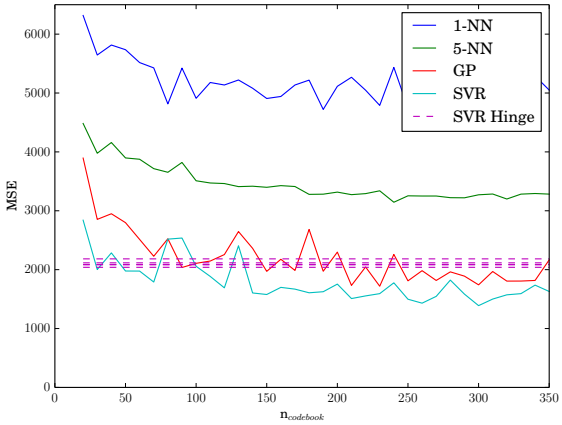

Figure 8: Graph showing the best performance for each category of codebook size. The 1-NN is very erratic, likely due to which shape context descriptor was found by the k-Means while creating the codebook. The 5-NN curve is smoother and can also estimate dates not represented in the training set. SVR Hinge includes the best parameter sets evaluated, and does not vary with codebook size. Note that above 140 in codebook size the proposed feature with SVR out-performs all other setups.

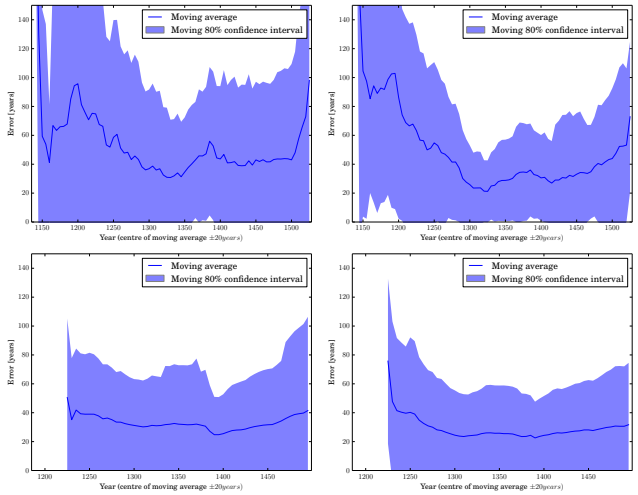

Figure 9: The moving error estimated at every point on the timeline in a window of ±20 years for 1-NN (upper left), 5-NN (upper right), Gaussian process regression (lower left) and SVR (lower right). This is an illustration for one run of the estimators i.e. the confidence interval is for the estimation errors on the individual charters and not because of stochastic effect in the estimators.

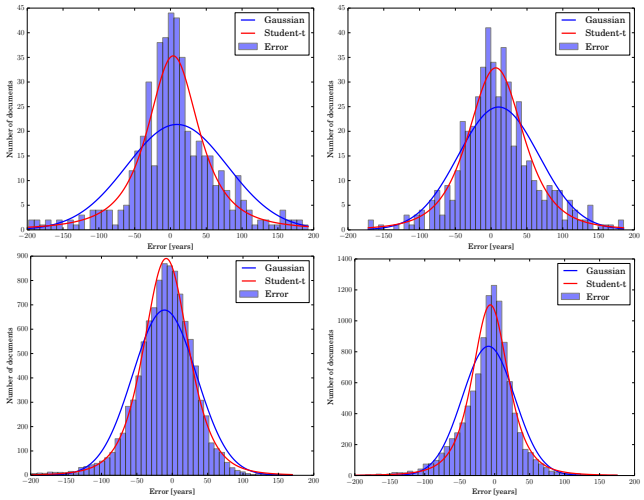

Figure 10: The distribution of error, in years, as a histogram for the proposed feature for 1-NN (upper left), 5-NN (upper right), Gaussian process regression (lower left) and SVR (lower right). The curves for the student-t and Gaussian distributions are the best fits for the data and are only included for reference to the shape of the distributions. Both reference curves are centered around 0, showing a bias in some of the estimators.

any value above 150 works well and makes the dimensionality far lower than for some comparable features.

The proposed method was robust to parameter variation but it would still be interesting to let a set of estimators vote for the best final dates, when inspecting the results from the different estimators that excel and fail at different points. In the future we would also like to evaluate our feature for writer identification and dating on other databases with both historical and modern material.

5. ACKNOWLEDGMENTS

This work was funded by Uppsala University and the Swedish Research Council (Dnr 2012-5743), within the q2b – From Quill 2 Bytes initiative, at Uppsala University.

6. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.

[2] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):679–698, 1986.

[3] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

[4] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963–2970, June 2010.

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531 vol. 2, June 2005.

[6] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *Image Processing, IEEE Transactions on*, 17(7):1178–1188, July 2008.

[7] S. He, P. Samara, J. Burgers, and L. Schomaker. Towards style-based dating of historical documents. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 265–270. IEEE, 2014.

[8] G. Louloudis, B. Gatos, N. Stamatopoulos, and A. Papandreou. Icdar 2013 competition on writer identification. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1397–1401, Aug 2013.

[9] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1832–1837, Nov 2005.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] C. E. Rasmussen. Gaussian processes for machine learning. MIT Press, 2006.

[12] M. Rusiñol and J. Lladós. Efficient logo retrieval through hashing shape context descriptors. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 215–222, New York, NY, USA, 2010. ACM.

[13] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of uppercase western script. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):787–798, June 2004.

[14] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1177–1178, New York, NY, USA, 2010. ACM.

[15] F. Shafait, D. Keysers, and T. M. Breuel. Efficient implementation of local adaptive thresholding techniques using integral images. In *Electronic Imaging 2008*, pages 681510–681510. International Society for Optics and Photonics, 2008.

[16] The GPy authors. Gpy: A gaussian process framework in python. http://github.com/SheffieldML/GPy, 2012–2014.

[17] F. Wahlberg, L. Martensson, and A. Brun. Scribal attribution using a novel 3-d quill-curvature feature histogram. In *14th International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, Crete, Greece, September 1-4, 2014*, pages 732–737, 2014.

[18] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn:

Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2126–2136, 2006.