Large scale continuous dating of medieval scribes using a combined image and language model

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

Lasse Mårtensson
Department of Business Studies
University of Gävle

Anders Brun
Department of Information Technology
Uppsala University

***Abstract*—Finding the production date of a pre-modern manuscript is commonly a long process in historical research, requiring days of work from highly specialised experts. In this paper, we present an automatic dating method based on modelling both the language and the image data.**

By creating a statistical model over the changes in the pen strokes and short character sequences in the transcribed text, a combination of multiple estimators give a distribution over the time line for each manuscript. We have evaluated our estimation scheme on the medieval charter collection "Svenskt Diplomatariums huvudkartotek" (SDHK), including more than 5300 transcribed charters from the period 1135 - 1509. Our system is capable of achieving a median absolute error of 12 years, where the only human input is a transcription of the charter text. Since reading and transcribing the text is a skill that many researchers and students have, compared to the more specialized skill of dating medieval manuscripts based on palaeographical expertise, we find our novel approach suitable for helping individual researchers to date collections of manuscript pages. For larger collections, transcriptions could also be collected using crowd sourcing.

I. Introduction

The dataset we have used for this paper was "Svenskt Diplomatariums Huvudkartotek" (SDHK). Some example images are shown in figure 6. The charters (in total over 10000) are from the Swedish medieval period and most are dated on the day. However, a significant portion of the charters have for multiple reasons lost their dates sometime during the last 500+ years. The dates were lost due to degradations, damage or that a part of the parchment was cut off for an unknown reason.

In [1], we developed a state-of-the-art pipeline for dating SDHK using only image based features. We draw heavily on that work in this paper for feature extraction and propose improvements to the estimators, more than halving the mean square error. We also explore and evaluate how adding a human to the estimation, by using transcriptions and rough estimates of the production dates, can improve the results. We find that the use of computer assisted dating, using our setup, can significantly lower the needed expertise and labour time of the user. The study we present here is a part of our project to return an estimation for the lost production dates in SDHK to the Swedish national archive (including expected estimation errors).

One of the main reasons for choosing the charters were that most are dated by the scribe, and thus represent a safe point of departure for the chronology. The most important usage for this method is to apply it to medieval manuscripts that are contemporary with SDHK, which as a rule are undated.

Traditional criteria for dating manuscripts are for instance palaeography, orthography, language forms, water marks (when the material is paper) etc. An overview over these criteria from the perspective of the medieval Nordic material is given by [2]. It should be noted that [2] is negative towards dating on palaeographic grounds on the stage of research when this work was produced. In the present investigation, we try to reestablish these criteria anew, and with new means. Rough palaeographic criteria such as the shape of certain characters (often used in older research) are probably too vague to allow for a more narrow dating. We focus instead on the more minute details in the formation of the script signs, details that in most cases have not been observed in traditional palaeographical research.

A. Previous work

In [3], [4], a multi step support vector regression (SVR) is proposed for increasing accuracy. This approach to regression was used in [5] for dating Dutch charters. In [5], [6], other features for dating are proposed and evaluated. A dutch charter collection very similar to SDHK but smaller (less than 1800 charters spanning the years $1295 - 1555$) was binned into 11 bins with charters form the years $\{1300 \pm 5, 1325 \pm 5, 1350 \pm 5, \cdots, 1550 \pm 5\}$. Thus the dating task was formulated as a classification problem with a number of classes corresponding to 10 year wide windows on the time line.

In [7], the authors propose a dating system for images of printed text based on convolutional neural networks (CNN). It also used the transcribed text for which the authors used OCR to get the transcriptions. Their data was binned by century $1600 - 1699, 1700 - 1799, 1800 - 1899, 1900 - 1999$. The final layer of the network was connected to the input from a non-normalized word level unigram model, including the sometimes very large category out-of-vocabulary (OOV). As is common when using CNNs for pattern recognition, the size of the training data had to be significant. The authors used a test set with a size roughly corresponding to 1% of the training set.

In [1], we proposed an image based approach for dating using a codebook of shape context descriptors [8], [9]. We

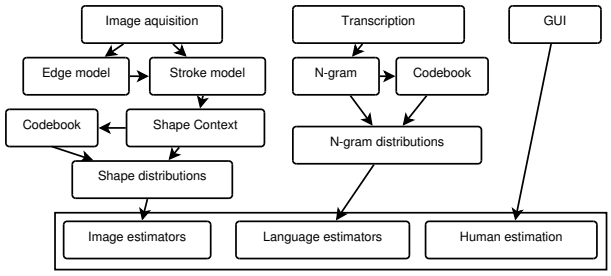

Fig. 1. A flowchart of the proposed system described in section II. The nodes at the top of the figure are the different types of input to the system. Information from them all were joint up in the bottom node where for each charter a distribution over production dates were estimated.

created a statistical model of changes to important stroke shapes over the time line. This was then evaluated using the aforementioned multi step SVR and using the same type of image features used in this paper. We concluded that using Gaussian processes (GP) for regression gave a slightly lower performance than using SVR. The size of the training set was only around 5% of the size of the test set. In the method described below, we have improved significantly on these estimator designs, mainly by combining multiple estimators and adding a language model.

II. Method

Our proposed system took as input the images, full transcriptions and an estimated date from a human. The SDHK charter collection contains over 5300 transcribed charters. Using the low resolution image (1,5 Mpix, 85% jpeg compression) and the transcribed text, we have trained several models based on Gaussian processes regression. The combined distributions over dates produced by these models were used as the final estimate. A flowchart of the full system is shown in figure 1. This feature extraction pipeline draws on the work we presented in [1] for the image features. The main differences from our earlier work on image based estimation is in the estimator design and in a new language model.

A. Image based features

The unsupervised feature learning for the images was performed on 1000 images (including the training set and filled up to 1000 using random images).

1) Edge distribution model: The Canny edge detection algorithm (proposed in [10]) needs two threshold parameters. Since the charter images differed in contrast and illumination, we estimated these parameters using a statistical model over the gradient magnitude of each charter image.

The Gaussian Mixture Model (GMM), shown in equation 1 (where x is a gradient magnitude), was fitted to the distribution over gradient magnitudes for each charter image. The GMM had two components to roughly estimate the low background magnitudes and the higher ink and parchment edges respectively. The magnitudes were created using Sobel operators and the l_2 norm.

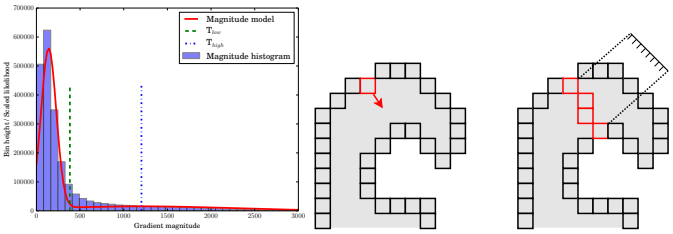

Fig. 2. **Left:** An illustration of how the threshold in section II-A1 were chosen. A Gaussian Mixture Model was fitted to the distribution of gradient magnitudes in each image. Thresholds for the edge detection are given by the value of the gradient magnitudes as fixed points in the mixture model (shown by the dotted vertical lines). **Right:** The stroke width (section II-A2) was measured from each edge point in the direction of the gradient over the ink. The euclidean distance to the end point was measured from the origin point to an edge point with an approximately opposite gradient direction.

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

The thresholds given to the Canny edge detector were set relative to the trained GMM. These fixed points, T_{high} and T_{low}, are defined as in equations 2 and 3. An illustration of the magnitude distribution, the model and the thresholds are shown in the left image of figure 2.

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

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

By using this automatic tuning of the Canny parameters, we were able to get a good edge detection on all of the 5000+ charter images without any human intervention. For the heterogeneous image data used here, estimating the parameters from the image was necessary for making a large scale date estimation possible.

2) Stroke width transform: In [11], the stroke width transform (SWT) was proposed. For each edge pixel, a line is drawn in the direction of the gradient (forward and backward) and the euclidean distance to the closest "opposite" edge are stored. "Opposite" is defined as an edge with a gradient direction with an angle opposite to the incoming line within a margin of $\pm\frac{\pi}{6}$ radians. An illustration is shown in the right image of figure 2.

We have used the SWT for de-noising the edge detection. It was assumed that all edge pixels belonging to a pen stroke have an opposite edge pixel within some scale dependent and predefined distance. If the distance was larger (or not found) the pixel was discarded. We set the cut off to 20 pixels to accommodate for possible scale differenced in the image collection. In figure 3, the output from the edge detection is shown on a part of a charter page.

3) Shape context: In [8], the shape context descriptor was proposed. This descriptor is a histogram over the pixels surrounding the center point in a patch of a set size (12 pixels in our case).

In a patch P, each edge pixel coordinate pair $(x_i, y_i) \in P$ was mapped to a log-polar coordinate system as $\theta_i =$

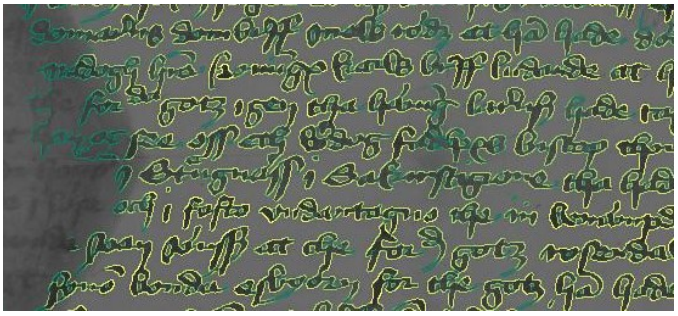

Fig. 3. A cropped gray scale charter image with the edges superimposed in colour. The edges were found using the Gaussian Mixture Model, Canny edge detection (section II-A1) and de-noised using the Stroke Width Transform (section II-A2). These were the edges used to create the shape context descriptor (section II-A3). The colour scheme goes from lower magnitudes in green to the stronger in yellow.

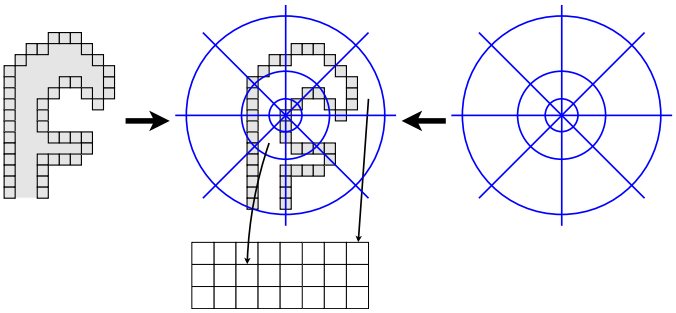

Fig. 4. The edge pixels (squares) in a patch are mapped to a histogram (shown by the blue stencil). This histogram was normalized, vectorized and used as the final descriptor belonging to the center point of the stencil.

$\text{atan2}(y_i, x_i)$ and $r_i = \log\sqrt{x_i^2 + y_i^2}$ into the new pair $(\theta_i, r_i) \in P$. Subtracting the gradient direction at the center point from each θ_i in a patch gave rotational invariance to the descriptor. The number of angle bins q, the number of log radius bins p and the patch side length N are model parameters. The process of creating a descriptor is also shown in figure 4. The logarithm function was used for giving more weight to the points closer to the center but also gives more scale invariance.

For each charter image, extracting this descriptor gave a 40 dimensional ($p = 5$, $q = 8$) point cloud representing the writer style.

4) Codebook training: To reduce the point cloud of descriptors from each charter to something more manageable, a reduction using a codebook was used. In [12], a mini batch k-means algorithm was presented that is suitable for large scale clustering. We created the codebook by clustering a point cloud of descriptors sampled from the charters in the training set using this.

By varying the codebook sizes ($N_{codebook} \in \{200, 240, 280, 320, 360\}$), we force the clustering to converge to different shapes for each run. Our procedure is very similar to the shapemes presented in [9], though they do not vary the codebook size. The different codebooks were later used to train different estimators.

After codebook training, new feature vectors were created using each codebook. Each charter's point cloud of descriptors were reduced to the relative frequency of each codebook entry.

B. Language based features

A problem in modelling language changes in pre-modern text is the lack of standardized spelling. In our source material, the spelling was not standardized in the Latin charters and even less so in the Swedish ones. In a model for dating, the changes in spelling over time could be an important indicator. To be able to create such a model the spelling would have to be so-called normalized (which can be complicated and is an active field of research in computational linguistics, see [13], [14]). A normalized text has the same spelling for all instances of the same word and variations are treated as misspellings. Transcribing then requires significant expert knowledge including understanding the subtleties of the text.

1) N-gram model: In this paper we have used an N-gram model with single characters as the smallest unit. In [15], this approach has successfully been applied to text categorization. We argue that spelling changes together with the changes in the language sounds, word choice and grammar of the languages being used by the scribes. Hence, doing statistics on the changes in frequency of single characters, character pairs and character triples have the potential to catch important changes over time in the use of a language.

An N-gram model has an order o defining the length of the character sequences used as features. We have used $o \in \{1, 2, 3\}$ in this paper. The transcribed text for each charter was tokenized into small (if $o > 1$, also overlapping) sequences of characters, including blank spaces. We pre-processed the transcribed text data by not letting the final text data include rare special characters (removed) or capitals (converted to lower case).

In some of the transcriptions, specific dates were mentioned. To account for this, we generated a separate n-gram model for all possible years (as text) in the time span using both Arabic and Roman numerals (e.g. text strings like "1234" and "MCCXXXIV" were used as input data). All tokens in this n-gram model of date strings were removed from the n-gram models of the charter text.

2) Feature space: The feature space spanned by the respective n-gram models were created by first finding all unique n-grams in the transcribed text for all charters. This unsupervised creation of the space gave us a dimensionality of around 40, 800 and 11000 for the 1, 2, and 3-gram models, respectively. The relative frequency of the keys in each charter were used as value for each dimension in the feature vectors. Hence, this was a bag-of-ngrams approach to the language modelling.

Using a 3-gram model was computationally costly due to the high dimensionality. To reduce the dimensionality of the 3-gram's 11000 dimensional model we performed PCA on the data, projecting it onto 500 and 1000 dimensional subspaces. This also improved the performance in terms of dating error of the 3-gram model greatly.

C. Estimator modeling

For modelling the changes in each feature space described above, we have used Gaussian process (GP) regression ([16]). A separate regression model was trained for each type of feature set (image/language based and feature parameters). The GP was chosen since it outputs a distribution over the output variables instead of a single estimate (as is the case with SVR). The feature space (often with a dimensionality in the hundreds) was mapped through the GP to a Gaussian distribution over the time line. For each charter, each feature set was given to the respective estimator giving multiple (one for each estimator) distributions over the time line. Each distribution was then treated as independent when combining them. The combined distribution over the time line was constructed as in equation 4 (where Φ is the set of estimators).

$$P_{combined} = \prod_{\Phi} P_{\phi}(t) \tag{4}$$

The chosen kernel function is defined between the feature vectors u and v as in equation 5 , a radial basis function (RBF) with automatic relevance determination (ARD). An ARD kernel includes weights (γ_i) for "stretching" the feature space in each dimension separately. The γ_i hyper parameters were allowed to be very small values that in effect removed some feature dimensions completely from the regression.

$$K_{ARD}(u, v) = \exp\left(-\sum_{i=0}^{|u|} \gamma_i (u_i - v_i)^2\right) \tag{5}$$

To determine the values of the hyper parameters, the hyper parameter space was optimized over using a gradient method. We have run several random restarts to ensure a good result and minimize the risk of getting stuck in a local maximum (the optimization maximizes the likelihood of the model, given the training data). The excellent implementation of Gaussian processes by [17] was used to create the estimators.

D. Human input

Crowd sourcing is today a popular approach to collect data from an digitized collection of historical material. Though this in some cases have been very successful, a dating task requires a very high level of expertise. Hence, it is unlikely that users could be found on the scale needed for most dating. We propose to use our pipeline as described above and combine the human estimate with that of the machine's. A less skilled used can often provide a rough estimate of the production date (e.g. within ± 100 years). Also, transcribing the data is a task that requires some knowledge of the character forms and development, but still a task that can be performed after a small amount of training, for instance by a student.

To evaluate the importance of a human expert, a human estimate and uncertainty was modelled as a distribution around the manually estimated date in the ground truth with a margin of uncertainty of x years (i.e. a tophat kernel with the width $2x$ or a Gaussian with $x = \sigma$). We want to show how accurate a human crowd sourced estimate needs to be before not benefiting from using a machine estimate for the date.

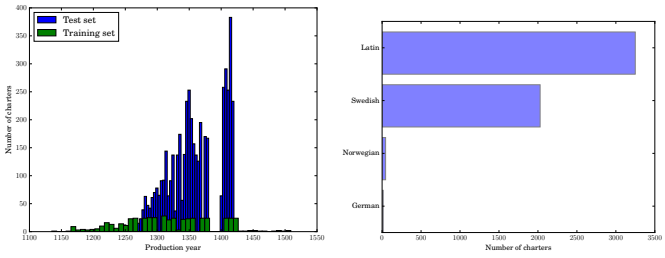

Fig. 5. **Left:** Histogram over the distribution of dates in the SDHK database. Here, the blue bars are the test set and the green bars the training set (for training set size of 500). **Right:** The number of charters written in the four most frequent languages. Note that the classes of languages do not refer to a static form (e.g. Swedish changes quite a lot during the studied period).

Fig. 6. Images showing examples from the collection "Svenskt Diplomatariums huvudkartotek" (SDHK), described in section III-A. The degradations on the manuscript pages include staining, holes and varying contrast. The quality was of "web" standard i.e. (in this case) 1.5 Mpix and 85% jpeg compression (with plenty of artifacts).

III. Evaluation

A. Svenskt diplomatariums huvudkartotek

The charter collection "Svenskt diplomatariums huvudkartotek" (SDHK) is a collection located at the Swedish national archive[1] and is the largest collection of charters from medieval Sweden. The full collection consists of over 11000 charters and more than 5300 of these are transcribed. In figure 5, a histogram over a number of charters is shown on a timeline. The collection spans several hundred years, from the $12th$ century to the $16th$. The random interest of some researchers determine which spans have been transcribed. Most of the charters are written in Latin (3200+) and old Swedish (2000+). In figure 5, the number of charters written in the four most frequent languages are shown. Some examples from the charter collection are shown in figure 6.

[1] http://riksarkivet.se/sdhk

TABLE I

EVALUATION METRICS FOR SOME OF THE ESTIMATOR COMBINATIONS. THE METRICS PRESENTED ARE THE $25th$ (P25), $50th$ (P50) AND $75th$ (P75) PERCENTILES OF THE ABSOLUTE ERROR TOGETHER WITH THE MEAN SQUARE ERROR (MSE). THE TRAINING SET SIZE ($N_{training}$) IS ALSO SHOWN FOR EACH ESTIMATOR.

Estimator	$N_{training}$	P25	P50	P75	MSE
Combined image & language	200	10	21	38	1152
Combined image & language	400	7	15	27	645
Single image	650	11	23	40	1422
Single language (3-gram)	650	9	19	33	1045
Combined image	650	8	17	30	810
Combined language	650	8	16	29	800
Combined image & language	650	6	13	24	525
Combined image & language	800	6	**12**	22	**462**

B. Evaluation metric

In equation 6 we show the mean square error (MSE) metric. We have chosen this metric since it punishes outliers significantly more than smaller errors. Small errors (e.g. ± 10 years) are of little importance in most applications and is considered insignificant by most researchers interested in historical material that we have come in contact with. Outliers, however rare, damages the trust in the system. Also, in a large scale implementation, even 1% outliers will be a large number.

The MSE metric is shown in equation 6 where X is the estimated dates and X^* the ground truth.

$$MSE(X, X^*) = \sum_{i=0}^{|X|} (X_i - X_i^*)^2 \tag{6}$$

We also present percentiles (e.g. 25th, 50th and 75th in table I) of the absolute error. This is to give a more intuitive metric of the performance of different types and sets of estimators. However, MSE work well for evaluating and ranking estimators.

C. Results

In table I, we show some evaluation results (expressed in the metrics described above). Note how the estimations became gradually better the more types of feature sets (i.e. sets extracted using different techniques) were included.

In figure 7, we plot the performance against the size of the training data. The full data set is 5300 charters (dated and transcribed in the ground truth). The union of the test and training set is always the full set for each chosen size of the training set (i.e. no data is withheld to keep the test set size equal for all sizes of the training set). It is often assumed that the accuracy of the model depends heavily on the size of the training set. Though this is not false, it is important to remember that the real factor is if the training data catches the variation in the full data set. By imposing a structure in our model through choices of both features and estimators, we have already constrained the final model significantly. At about 500 charters in the training set, our model had caught enough variance for the MSE to level out.

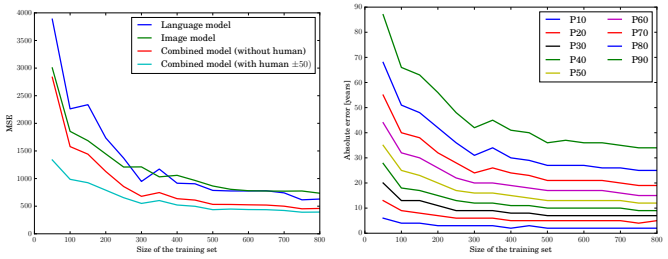

Fig. 7. Performance (y-axis) plotted against the size of the training data set (x-axis). **Left:** The plot shows the MSE for the estimators based on the transcribed text, the images and the combined estimator using both. **Right:** The decile limits of the absolute error for the combined estimator is shown. Note that with a training set size of 500 (i.e. $< 10\%$ of the full data set), the performance does not increase much by using more data.

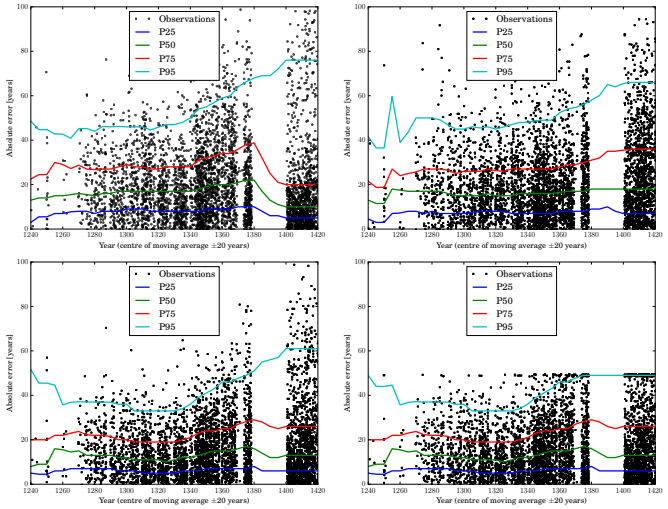

Fig. 8. All sub-figures show the moving error over the time line. The coloured lines are the 25th, 50th, 75th and 95th percentiles of absolute error in the ± 20 year area around each year on the time line (x-axis). **Upper left:** Estimator using only language modelling **Upper right:** Estimator using only image modelling **Lower left:** Combined estimator for both language and image data **Lower right:** The Combined model with a human expert modelled as a uniform distribution around the true date ± 50 years.

In figure 8, moving error estimates are shown. Note that using a human (lower right) sets a cap on the outliers. This has a large effect on MSE but not on lower percentile errors.

In figure 9, the distribution of the estimation errors can be compared while showing data for "Swedish" and "Latin" separately. We show this because the language used (and not only the use of a language) was correlated with time.

Our system was designed for large scale production date estimation. However, adding a human expert will likely improve the final production date estimate. In figure 10, the performance of the proposed system is compared against a modelled human expert. We aim to answer how accurate a human has to be to improve on the automatic system. The human is modelled as a uniform distribution with varying width or a Gaussian distribution with varying standard deviation. With this comparison we want to show to which accuracy a human must perform to be comparable to the proposed automatic

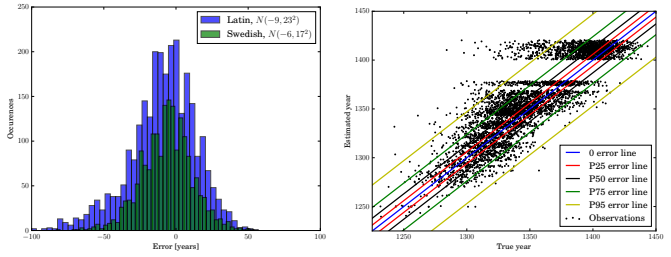

Fig. 9. Visualizations of the error using $N_{training} = 650$. **Left:** Histograms over the error in estimations for the groups of charters in Latin and Swedish. The error distributions look very similar though the bias and spread differ slightly. Note that the parameters of the respective distributions are included in the legend text (treating the distributions as Gaussian). **Right:** A scatter plot over the errors with percentiles of error as diagonal lines through the plot.

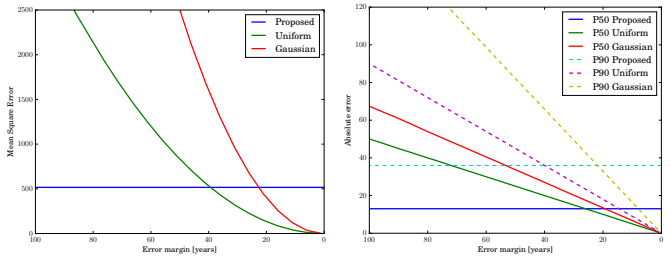

Fig. 10. Error comparisons to determine the performance demands on a human expert to be comparable to the proposed system while varying the margin m. The human error is modeled as a uniform distribution $U(-m, m)$ and a Gaussian distribution $N(0, m^2)$. **Left:** The MSE for the proposed system ($N_{training} = 650$) compared to a model of a human expert. **Right:** The 50th and 90th percentiles of the absolute error for the proposed system compared to a model of a human expert.

estimation.

IV. Conclusions

In this paper, we have shown that image- and language based models can be used for dating of medieval charters from a large and heterogeneous collection. The combination of several Gaussian process estimators improve the result even further. In particular, our novel model based on transcribed text is shown to greatly improve the estimation of production dates, despite being based on relatively simple character n-grams. Our results indicate that the image and text data estimators bring different information to the final estimate. The transcriptions that are necessary for creating the language models are fairly easy to obtain, e.g. from trained students or researchers that do not have to be experts in paleography. This makes our approach suitable for larger collections, using crowd sourcing, which is actually the case for SDHK.

As the results in figure 10 show, human experts have to be very accurate in their estimation to be better than our machine dating based on image- and language features, on average. However, using a human for finding outliers in the estimation can also be very beneficial, as show in figure 8. This is a realistic scenario when a relatively skilled user, albeit not an expert paleographer, wants to date a particular document. In summary, this means that the different variants of our method make the demand on expertise much smaller for a human using our computer assisted dating system. And in the cases where image information, transcriptions and a preliminary user estimate of the production date is available, the combined estimate will be the best.

Acknowledgment

We would like to thank the Swedish National Archive and Sara Risberg for their generous help in providing both images, metadata and expertise. Also, without the enormous contribution by Per-Axel Wiktorsson on the SDHK charter collection, this study would not have been possible.

References

[1] F. Wahlberg, L. Mårtensson, and A. Brun, "Large scale style based dating of medieval manuscripts," in *The 3rd International Workshop on Historical Document Imaging and Processing (HIP15)*, 2015.

[2] P. A. "om, *Senmedeltida svenska lagbcker 136 lands- och stadslagshandskrifter, dateringar och dateringsproblem*. Acta Universitatis Stockholmiensis, 2003.

[3] 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*, vol. 17, no. 7, pp. 1178–1188, July 2008.

[4] 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*, vol. 2, 2006, pp. 2126–2136.

[5] 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*. IEEE, 2014, pp. 265–270.

[6] S. He and L. Schomaker, "A polar stroke descriptor for classification of historical documents," in *International Conference on Document Analysis and Recognition*. IEEE, 2015.

[7] Y. Li, D. Genzel, Y. Fujii, and A. Popat C., "Publication date estimation for printed historical documents using convolutional neural networks," in *The 3rd International Workshop on Historical Document Imaging and Processing (HIP15)*, 2015.

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 4, pp. 509–522, 2002.

[9] G. Mori, S. Belongie, and J. Malik, "Efficient shape matching using shape contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 11, pp. 1832–1837, Nov 2005.

[10] J. Canny, "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 679–698, 1986.

[11] 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*, June 2010, pp. 2963–2970.

[12] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1177–1178. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772862

[13] E. Pettersson, B. Megyesi, and J. Nivre, "A multilingual evaluation of three spelling normalisation methods for historical text," *Proceedings of LaTeCH*, pp. 32–41, 2014.

[14] M. Piotrowski, "Natural language processing for historical texts," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 2, pp. 1–157, 2012. [Online]. Available: http://dx.doi.org/10.2200/S00436ED1V01Y201207HLT017

[15] W. B. Cavnar, J. M. Trenkle *et al.*, "N-gram-based text categorization," *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175.

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

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