Machine Dating of Handwritten Manuscripts

Utpal Garain[1] **S. K Parui**[1] **T. Paquet**[2] **L. Heutte**[2]

1. *Computer Vision & Pattern Recognition Unit*
   Indian Statistical Institute,
   203, B. T. Road, Kolkata 700108, India
   Email: {utpal,swapan}@isical.ac.in

2. *Laboratoire LITIS EA 4051,*
   UFR des Sciences, Université de Rouen,
   76800 Saint-Etienne du Rouvray, France
   Email: {Thierry.Paquet, Laurent.Heutte}@univ-rouen.fr

Abstract

This paper presents a pioneering study on automatic dating of handwritten manuscripts. Analysis of handwriting style forms the core of the dating method. Initially, it is hypothesized that a manuscript can be dated, to a certain level of accuracy, by looking at the way it is written. The hypothesis is then verified with real samples of known dates. A general framework is proposed for machine dating of handwritten manuscripts. Experiments on a database containing manuscripts of Gustave Flaubert (1821-1880), the famous French novelist reports about 62% accuracy when manuscripts are dated within a range of five calendar years with respect to their exact year of writing.

1. Introduction

Handwritten manuscripts have been playing a crucial role in recoding and transferring knowledge from the ancient ages. Many of these documents have been produced in printed and translated editions, but any new evidence for the historian, the social scientist or the genealogist must be acquired through the reading of the original material. The quantity of these historical manuscripts stored in archives, libraries and private collections is vast and researchers in the field of archaeology, document decoding, literary science, or paleography and diplomacy, etc. have engaged themselves in different kinds of studies focused on these manuscripts.

Among these groups one section considers dating of manuscripts to provide an answer to questions like when was *Beowulf* [1] written, or in which year did James Joyce draft a particular chapter of *Ulysses*, or when did Abraham Lincoln write a particular undated letter, etc. To answer these questions one may think of age determination by radiocarbon method [1, 2] that has been a standard practice in science. However, the radiocarbon dating of manuscripts is very rare as this technique is not only very expensive but also provides a wide range of calendar years in which a particular manuscript could have been written. Sample type, sample size, sample handling (samples undergoing radiocarbon tests get destroyed), etc. often discourage the application of this method for dating manuscripts.

On the other hand, several palaeographical techniques [3] have been available in the literature for dating of handwritten manuscripts. Methods consider one (or many in combination) of several techniques based on (i) style of handwriting, (ii) paper watermark (often dated or datable), (iii) examination of the content of the text often gives clues, (iv) the binding (if contemporary), etc. In contrast to radiocarbon dating, success of a palaeographical method often depends on human expertise. Many views that such expert opinions based on some scientific techniques are inadmissible unless the technique is established based on testing, peer review, error rates, and acceptability. The proposed work is motivated by this scientific need. This study attempts to establish some of the palaeographical practices with scientific rigor and at the same time it endeavors to explore the possibility of involving automatic means for dating manuscripts. The later effort is essentially aimed at helping the human experts in making final decisions while dating a manuscript in question.

Experts have often viewed that a manuscript can be dated, to a certain level of accuracy, simply by looking at the way it is written. Handwriting is a product of human culture and as such it is always developing. Therefore, paleographers hypothesize that differences in handwriting are bound to appear within a period of time and based on this hypothesis they often attempt to date manuscripts.

This paper makes use of this assumption and incorporates statistical techniques to validate this hypothesis using real samples with known dates. Accuracy of the proposed method in dating

[1] *Beowulf* (c. 700-1000 A.D.) is a heroic epic poem.

manuscripts is then verified. The present experiment has restricted itself to a set of manuscripts all written by a single individual over a wide span of calender years and then a general framework is proposed and tested on this manuscript set to answer two types of queries: (i) whether a pair of manuscripts are written at the same time or within a certain period of time and (ii) whether a manuscript of unknown date can be dated. The rest of the paper discusses about the hypothesis, statistical methods used for its validation and experiments checking the usefulness of the proposed approach for dating manuscripts with real samples.

2. Construction of the hypothesis

Consider the situation when we have a collection of handwritten manuscripts written by a writer over K successive years. Manuscripts written within a certain time period are grouped together. Let w ($1 \le w \le \lfloor K/2 \rfloor - 1$) be the span of this time period and two manuscripts that differ in number of years less than w fall in the same group. Following this way, manuscripts written in successive K years will give G (= K-w+1) number of groups.

Let g_k be an individual group and a manuscript may belong to more than one (maximum up to w) groups as groups are not disjoint. Let x_i^k denote the i-th (i = 1, 2, …, l_k) manuscript in the k-th group g_k , and l_k ($\ge w$) denote the number of manuscripts in that group. We assume that manuscripts within each group represent homogeneity in the sense that they belong to the same writing style. On the other hand, manuscripts written several years apart are assumed to represent heterogeneous pairs in the sense that they belong to different handwriting styles.

Similarity in writing style between a pair is measured by employing a method described in details in [4]. In this approach, the image of a handwritten manuscript initially goes through a pre-processing phase. A handwritten manuscript is processed to find connected components that are then segmented into graphemes. Graphemes are then clustered using an unsupervised classification step to define a set of features pertaining to that manuscript. Let M be the feature vector $M = \{m_1, m_2, \ldots, m_n\}$ common to two analyzed pages written in calendar years Y_1 and Y_2. Let Y be the set of these two calendar years, $Y = \{Y_1, Y_2\}$.

A relation between Y and M is established by computing mutual information between them:

$$I_m(M,Y) = H(M) - H(M \mid Y) \quad (1)$$

where $H(M)$ indicates the Shannon entropy [5]

$$H(M) = \sum_{i=1}^{|M|} P(m_i) H(M = m_i) \quad (2)$$

and $H(M|Y)$ indicates the conditional entropy. Mutual information in (1) can be equivalently expressed as

$$I_m(M,Y) = H(M) + H(Y) - H(M,Y) \quad (3)$$

where $H(M, Y)$ is the joint entropy of M and Y.

Mutual information (*MI*) thus measures independence between the set of two years and the set of features. Low *MI* value indicates the same/nearly the same value for the calendar years Y_1 and Y_2 whereas high value indicates a strong dissimilarity between them. To assess the role of this criterion, an experiment is carried out on the handwriting of Gustave Flaubert (1821-1880) the famous French novelist. A large number of manuscripts (including the manuscript of *Madame Bovary*) written by Flaubert are available with the library of Rouen, France. A portion of this collection has been used here and the database (termed as *Flaubert database* in this paper) consists of one or more manuscripts for each year starting from 1846 to 1880. More precisely, the database used for this experiment consists of correspondences of Flaubert each having as the date of writing.

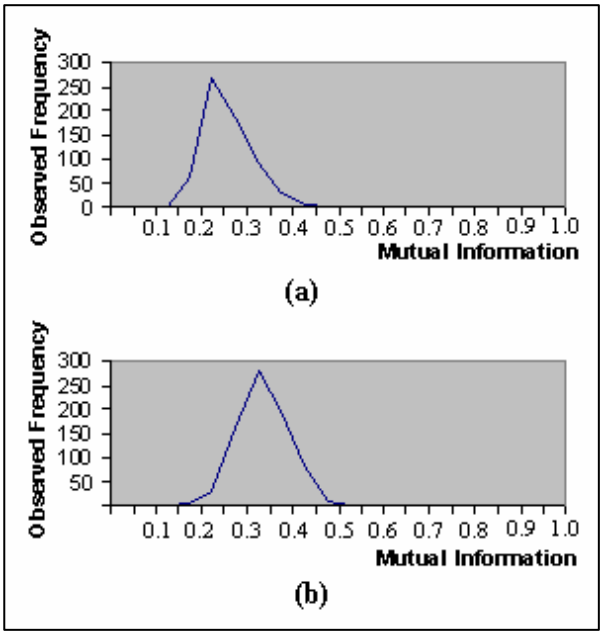

Figure 1. Distributions of the mutual information criterion for (a) homogeneous and (b) heterogeneous samples.

3. Hypothesis test

Let g_k denote a group of manuscripts all written within a span of w years (in this present experiment, w is set to 5) and l_k be the number of manuscripts in this group. Then there are $C_2^{l_k}$ mutual information (*MI*) values one for each of the $C_2^{l_k}$ pairs that can be formed by the manuscripts in g_k. Since there are G such groups resulting in $N\ (= \sum_{k=1}^{G} C_2^{l_k})$ number of *MI* values. This forms the space for homogeneous samples. Distribution (f) of these homogenous cases (in total, 636 number of such homogeneous pairs are in the database when w is set to 5) in Flaubert database is shown in Fig. 1 (a). Note that an *MI* value can take any real value in [0, 1]. This range is divided into 20 equal intervals (q_1, q_2, …, q_{20}) each of width 0.05, i.e. 0.0<= q_1 <=0.05, 0.05< q_2 <=0.1, …, 0.95< q_{20} <=1.0. Frequency of samples in each of these intervals is computed to find the distribution.

Heterogeneous samples are formed as follows. The groups are initially arranged in the ascending order of the earliest dated manuscript in each group. Let g_1, g_2, …, g_G be this ordering. Next, manuscripts in two groups g_i and g_j are considered where i = 1, 2, …, G/2 and $j = \lfloor G/2 \rfloor + i$ (for G even) or $\lfloor G/2 \rfloor + i + 1$ (for G odd). This results in $\lfloor G/2 \rfloor$ number of groups $g_{i,j}$ (= g_i x g_j) where each such group considers mutual information between a pair of manuscripts in which one comes from g_i and the other from g_j. Therefore, each of these $g_{i,j}$ contains $|g_i|$ x $|g_j|$ *MI* values. As the manuscripts in any pair of these groups differ in a minimum of w (when $w = \lfloor K/2 \rfloor - 1$) years and a maximum of $\lfloor G/2 \rfloor$ years (when w = 1), these $\lfloor G/2 \rfloor$ groups are assumed to represent heterogeneous samples. Distribution (f') of these heterogeneous cases (in total, 720 number of such heterogeneous pairs are in the database when w is set to 5) in Flaubert database is shown in Fig. 1 (b).

3.1. Multinomial Chi-square test

Samples of each group g_i are selected randomly to divide the group into two equal halves. First half of each group forms a new distribution (f_1) and the second half results in another distribution (f_2). These distributions on the Flaubert database are shown in Figs. 2 (a) and (b). From the observation of these two distributions (both presenting homogeneous cases) in Fig. 2 and earlier distributions in Fig. 1 (homogeneous vs. heterogeneous) it seems that mutual information could provide a quantitative criterion for the task of

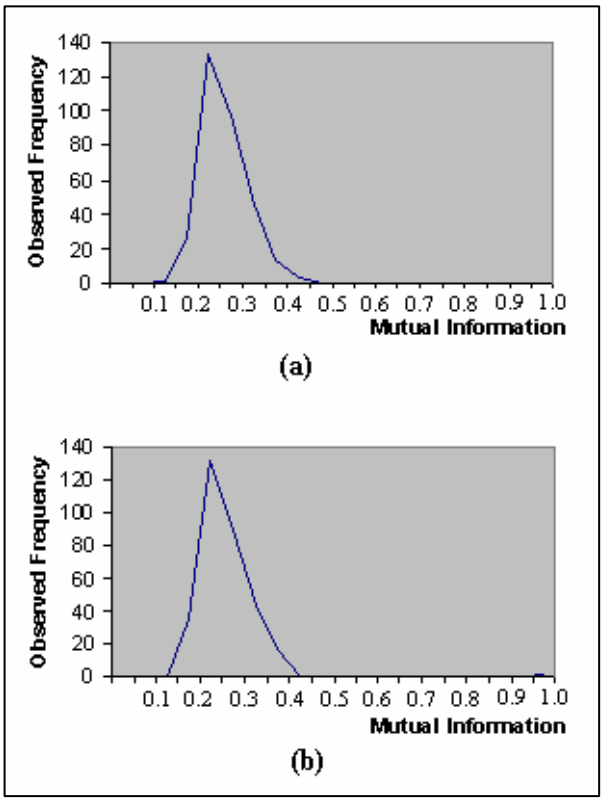

Figure 2. Distributions of the mutual information criterion for two sets of homogeneous samples.

dating manuscripts. This observation is further verified by two experiments first of which is given by:

$$H_0: f_1 = f_2 \tag{4}$$

where H_0 is the null (or default) hypothesis. To test this hypothesis a multinomial chi-square test [6] is conducted.

Let E_{ij} be the expected number of observations. Under H_0: $E_{ij} = N_i * O_j / N$, in which O_j are the total number of observations of categories j (j = 1, …, J, in this experiment J = 20), N_i the sizes of samples i (i = 1, …, I; in this experiment I = 2, i.e. f_1 and f_2). Note that f_1 and f_2 each contains N/2 number of samples and hence, $N_1 = N_2 = N/2$. Finally, the test parameter χ^2 is computed as follows:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{20} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

which follows a Chi-square distribution with (J-1)*(I-1) degrees of freedom. Note that O_{ij} are the number of actual observations in the j-th category (or interval) for samples i.

Experiment on Flaubert database gives a value near 15 for χ^2 as defined in (5). As samples are randomly selected to divide each g_i, we observe slight variations in distributions f_1 and f_2 each time the randomization

algorithm is initialized with different seeds. Ten runs with ten different initial seed values for randomization were conducted to compute the value of χ^2 in (5) and it is noted that the values are in the range [9, 20]. The degrees of freedom (df) in this experiment is 19 and when the chi-square table is consulted significance level is 0.10 at χ^2 = 27.2036 for df=19. If the probability (Q) of the result being due to chance is computed[2], we obtain Q = 0.3945 for χ^2 = 20 and Q = 0.9734 for χ^2 = 9 corresponding to df = 19. This strongly attests the assumption of the null hypothesis in (4).

The null hypothesis of the second experiment is given by

$$H_0: f = f' \tag{6}$$

The test is conducted in a similar manner as described before. A high value (i.e. 339.4391) is obtained as measure of χ^2 between f and f'. When Chi Square table is consulted for degrees of freedom equals to 19, it is found that the significance level comes down to 0.005 at χ^2 = 38.5821 and when the probability (Q) of the result being due to chance is computed, we obtain Q = 0 for χ^2 = 339.4391. Hence, H_0 in (6) cannot be accepted.

4. Accuracy in dating manuscripts

After the hypotheses are statistically verified in the preceding section, this section seeks answers to two types of questions: with what accuracy (i) a pair of manuscripts can be classified as homogeneous (i.e. both of them are written within a period of w years) or heterogeneous? (ii) A manuscript of an unknown date can be dated correctly?

To answer the first query, a Beta distribution [6] of the mutual information is assumed. This assumption is based on the fact that (i) mutual information lies between 0 and 1 and (ii) a Beta distribution can have a shape that is not necessarily symmetric.

Beta distribution is defined as,

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{7}$$

where $0 < x < 1$ is a Beta random variable (here mutual information), parameters α and β are greater than zero, and B is the Beta function. The method-of-moments estimates of the parameters are

$$\alpha = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{v} - 1\right) \text{ and } \beta = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{v} - 1\right) \tag{8}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the sample mean and $v = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$ be the sample variance.

Following (6) α and β are computed from f (figure 1(a)) i.e. the distribution of homogeneous samples and α = 15.064 and β = 43.189 are obtained for the Flaubert database.

The beta function $B(\alpha, \beta)$ is computed as

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{9}$$

where Γ is the gamma function.

In practice, an MI (mutual information) value is obtained from a pair of manuscripts and using this value the pair is to be marked homogeneous or not. This task requires a threshold value against which this decision can be taken. To determine that threshold (say, y) a function $g(y)$ is defined as follows to give the coverage (or confidence) of such a decision for a certain value of y.

$$g(y) = \int_0^y f(x;\alpha,\beta)\, dx \tag{10}$$

Two values of y, y_1 and y_2 are computed so that $g(y_1)$ = 0.95 and $g(y_2)$ = 0.99. The experiment on the Flaubert database gives y_1 =0.36 and y_2 =0.46. This indicates if we set the threshold at 0.36 then nearly 95% of the homogeneous manuscript pairs will be classified correctly as homogeneous. On the other hand, if the threshold is set at 0.46 then about 99% of them will be properly classified.

However, $g(y)$ does not give accuracy in classifying heterogeneous pairs. Therefore, mutual information values for these pairs (known heterogeneous) are checked against y_1 and y_2 to compute the correct classification accuracy. In the Flaubert database, out of 720 pairs of heterogeneous samples, classification errors are 35 and 192 at y_1 and y_2, respectively. Table 1 summarizes the accuracies in classifying a pair of manuscripts as homogeneous or heterogeneous.

Table 1. Accuracies for classification of a pair of manuscripts as homogeneous/heterogeneous

Type	#Samples	At $y = y_1$	At $y = y_2$
Homogeneous	636	0.955	0.993
Heterogeneous	720	0.951	0.733

To provide an answer to the second query i.e. accuracy of proper time stamping of a manuscript of

[2] To calculate Q from χ^2 and df, a calculator is available at http://bavard.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

unknown date, we assume that datation of a manuscript is appropriate if this falls within a certain neighborhood of the year in which the manuscript was exactly written. We define this neighborhood by using w. A manuscript written in the year k is dated accurately if machine returns a value in [k-(w-1), k+(w-1)] to date the manuscript. The rationale behind defining this neighborhood lies in the assumption of homogeneity of manuscripts as discussed in section 2. With respect to a manuscript written in the year k, manuscripts written in the year k-(w-1) and k+(w-1) are considered as the farthest homogeneous manuscripts on the left and right side of the time scale, respectively.

In doing so, manuscripts written in each year define a set of features common to all the same-year manuscripts: $F = \{f_1, f_2, \ldots, f_n\}$. Let F_i denote the feature vector for year i. A target manuscript (say, X) is dated by measuring mutual information between F_i and X for all i by following the equation in (1) and is dated as written in year j if

$$j = \arg \min_i I_m(F_i, X) \tag{11}$$

Thirty four manuscripts of known dates were chosen from the Flaubert database and accuracy of automatic datation scheme as described above is computed. It's verified that when w is set to 5, accuracy of correctly dating the manuscripts is about 62% and this rate decreases with decrease in w. Figure 3 shows the effect of w on datation accuracy.

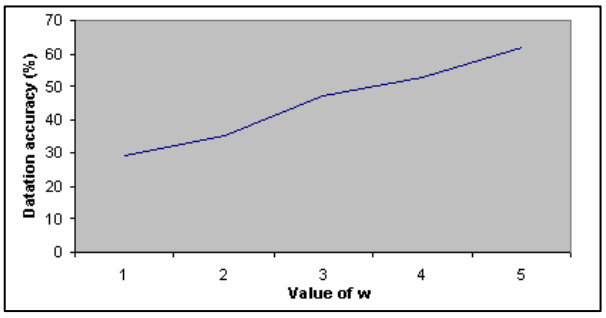

Figure 3. Accuracy for automatic datation of manuscripts.

5. Conclusions

Automatic dating of handwritten manuscripts is studied in this paper. Like paleographers it is hypothesized that differences in handwriting are bound to appear within a period of time and based on this hypothesis one may attempt to date manuscripts. Later, this hypothesis is statistically verified using real samples. A general framework is proposed to answer two types of queries: (i) whether a pair of manuscripts is written at the same time or within a certain period of time and (ii) whether a manuscript of unknown date can be dated. Experiments on a set of manuscripts written by Gustave Flaubert, the famous French novelist, reveal that with more than 95% confidence we can provide answer to the first query. On the other hand, manuscripts of unknown dates were dated with an accuracy of about 62%.

The present experiment attempts to establish the paleographical practices especially involvement of handwriting styles in dating manuscripts with more scientific rigor than it was before. The current study, therefore, would provide a useful assistance to the experts who are involved in this business.

Future extension of this study includes incorporation of newer methods of handwriting analysis to improve the accuracy of automatic datation of manuscripts. At the same time, experiments on other datasets are needed to attest the potentiality of the proposed approach. Instead of assuming the beta distribution of mutual information criterion, consideration of a normal distribution would be an additional aspect of the future study. So far experiment has been restricted to analysis of handwriting of an individual. This would be extended in future to consider handwriting styles pertaining to multiple writers. The proposed framework with further extension could provide a useful tool for dating manuscripts of ancient ages.

6. References

[1] Libby, W.F., Anderson, E.C., and Arnold, J.R., "Age Determination By Radiocarbon Content: World-Wide Assay of Natural Radiocarbons", Science, Volume 109, pp. 227-228, 1949.

[2] Taylor, R.E., "Radiocarbon Dating: An Archaeological Perspective," 1987, Academic Press, Inc.: Orlando (FL), pp. 169-170.

[3] Watson, A.G., "Catalogue of Dated and Datable Manuscripts c.700 - 1600 in the Department of Manuscripts," The British Library 2 volumes, London: British Museum Publications, 1979.

[4] Bensefia, A., Paquet, T., and Heutte, L., "A writer identification and verification system," Pattern Recognition Letters (PRL), 26, 2080-2092, 2005.

[5] Shannon, C., "The mathematical theory of communication," Bell System Tech. J. 27, 379-423, 1984.

[6] Hoggs, R.V., Tanis, E.A., "Probability and Statistical Inference," Prentice Hall, 5 edition, NJ, USA, 1996.