Volume 30, No. 1, 1994 Januar... Februar...

INFORMATION PROCESSING & MANAGEMENT

An International Journal

INFORMATION PROCESSING & MANAGEMENT
-An International Journal-
(Incorporating *INFORMATION TECHNOLOGY)*

Volume 30, Number 1 **1994**

Contents

(Continued on inside back cover)

INDEXED IN Cum. Cont.", Research Alert®, Biosir Data, Cam. *Sci.* Abstr, Chem. Abstr., CompuMath Cit. Ind., Health Serv. Admin., Curr. Cont./Soc. & Beh. Sci., Curr. Cont./Eng. Tech. & Applied Sci., Comput. Cont., Curr. Tech. Ind., ERIC Clear on Inform. Res., Eng. Ind., Ergon Abstr., Info. Sci. Abstr., LISA, Leg. Cont., Lib. Sci. Abstr., Lib. Lit., Manage Cont., PASCAL-CNRS Data, PIRA, Soc. Sci. Cit. Index®, SSSA/CISA/ECA/ISMEC.

PRINTED IN GREAT BRITAIN by A. WHEATON & CO. LTD., EXETER.

ISSN 0306-4573
IPMADK 30(1) 1-162 (1994)
244

Information Processing & Management Vol. 30, No. 1, pp. 141–150, 1994
Printed in Great Britain.
0306-4573/94 $6.00 + .00

DISCRIMINATION OF AUTHORSHIP USING VISUALIZATION

BRADLEY KJELL
Computer Science Department, Central Connecticut State University.
New Britain, CT 06050, U.S.A.

W. ADDISON WOODS*
United States Army Student Detachment, Fort Benjamin Harrison,
Indianapolis. IN 46216, U.S.A.

and

OPHIR FRIEDER
Computer Science Department, Center for Image Analysis.
George Mason University, Fairfax, VA 22030, U.S.A.

(Received 14 September, 1992: accepted in final form 4 February, 1993)

Abstract – Visualization techniques help organize the vast amount of data generated in computational studies of literary style. These techniques are demonstrated by showing two-dimensional representations of the style of the authors of *The Federalist Papers*. The techniques are used to determine the authorship of the 12 unattributed papers. The authorship assigned to these papers is consistent with that found in other studies.

1. INTRODUCTION

Computers frequently have been used to characterize literary style by the values of parameters extracted from text. These characterizations have solved questions of disputed authorship, have indicted changes in an author's style with time, and have shown the fluctuations in style with changes in the mood of a work. Most models of style have used easily quantifiable features. These features largely fall into three groups: word and sentence length features, vocabulary features, and syntactic features, as seen in Hockey (1980) and Holmes (1985).

Early studies of style assumed that works from different authors would exhibit different frequency distributions for word and sentence length. Mendenhall (1887) used word-length distribution statistics to study the question of the authorship of the Shakespearean plays. Mosteller and Wallace (1964) use sentence length and vocabulary to solve the problem of authorship in The *Federalist Papers*. Other researchers use the distribution of function words, such as articles and connectives (Kenny, 1986), or the distribution within sentences of words used only once in the text. Often, a combination of several such features is used (Stratil & Oakley, 1987).

Many of the early studies were done on mainframe computers using small, laboriously keypunched samples of text. With modern text scanners and computing equipment, it is now possible to obtain the complete text of works being studied and to use a rich set of features. Along with these new capabilities comes the problem of organizing the potentially vast amount of data and choosing the most incisive features for describing literary style. Visualization techniques help in these problems.

2. LETTER-TUPLE FREQUENCY STATISTICS

Much of the style of an author is contained in the statistics of N-tuples of letters extracted from a sample of the author's work (Bennett, 1976; Hayes, 1983; Kjell, 1985;

*Major Woods is an officer in the United States Army currently serving on active duty as a full-time student. Since September 1992, he has been assigned to the United States Army Student Detachment with duty at George Mason University, School of Information Technology and Engineering, as a doctoral student.

Requests for reprints of this article should be submitted to Prof. Bradley Kjell, Department of Computer Science, Central Connecticut State University, 1615 Stanley Street, New Britain, CT 06050.

Tankard, 1986). This method will be illustrated using *The Federalist Papers*. These are 85 papers published anonymously in 1787-1788 by Alexander Hamilton, John Jay, and James Madison, discussing aspects of the Constitution. As Rossiter (1961) explains

> This mask of anonymity, put on by the authors for sound political purposes, made it possible for Hamilton. in a note written just before his death and discovered just after, to lay claim to a full 63 numbers of *The Federalist*, some of which very plainly belonged to Madison.

Specifically, the authorship of 12 documents (numbers 49-58, 62, and 63) have historically been debated, but is now generally attributed to Madison (Rossiter, 1961).

Various methods, such as those discussed earlier, have since been used to determine Madison's authorship of these documents. Here, the method of 2-tuple and 3-tuple frequency will be applied to analyze these 12 historically disputed papers to compare the results of this technique to those of previous methods. The ASCII text for this experiment was obtained from *Project Gutenburg* (Project Gutenburg Association, Illinois Benedictine College: ftp mrcnext.cso.uiuc.edu). The numbering of papers and the attribution of authorship, when known, follows that text,

Two prototype texts were prepared by concatenating all of Madison's papers into one file (papers 10, 14, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, and 48), and concatenating a selection of Hamilton's papers into another file (papers 1, 6, 7, 8, 9, I I, 12, 13, IS, 17, 27, 68, 70.71, 72, 73, 74, 75, 76, and 77), such that the files were approximately the same length (237,000 characters). Both early and late Hamilton papers were picked, so that any change in style would be included in the prototype. A third prototype text was prepared by concatenating all of the documents of unknown authorship (papers 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 62, and 63) into *one* file of approximately 135,000 bytes in length. This third prototype represents the collection of text of disputed authorship, but believed to be written as a collection by either Madison or Hamilton (Rossiter, 1961).

First, we will discuss the details and experimental results of two-letter tuples, followed by three-letter tuple results, omitting the redundant computational details. Each prototype was processed by counting all occurrences of two-letter tuples. Only characters in the range a . . . z were used, upper case was converted to lower case, and punctuation, including spaces. was ignored. Tuples overlap, so each letter was a member of two tuples. Tuple *th* is the most frequent, with a relative frequency of about 0.026, and many tuples such as *qq* occur with a frequency of zero. The representation for each prototype text, or any of the individual documents, is a vector of 26^2 features (where most values are zero). Let X be the vector for Hamilton, and Y be the vector for Madison. The cosine of the angle between two feature vectors is

$$\cos(\theta) = \frac{\mathbf{X}'\mathbf{Y}}{\|\mathbf{X}\|\,\|\mathbf{Y}\|},$$

where X'Y is the inner product of the vectors and $\|\mathbf{X}\|$ is the Euclidean norm of the vector. This is the cosine similarity measure commonly used in information retrieval (Salton &McGill, 1983). Co-linear vectors will have a cosine of 1.000; dissimilar vectors will have a smaller cosine. Each of the 85 papers may be compared with the two prototypes. A feature vector for each paper is created as above, and the cosine similarity measure is computed between that vector and each of the prototype vectors.

These results are summarized in Table I and are listed in detail in Table 3, found in the appendix. The row label *m10* in Table 3 designates the tenth Federalist paper, which was written by Madison. The label j02 designates a paper written by Jay; *h01* designates a paper written by Hamilton; *b/8* designates joint authorship between Hamilton and Madison: and *u49* designates uncertain authorship (either Hamilton or Madison). Additionally, we compare the Madison and Hamilton prototypes to each other and observe that a prototype document has a cosine similarity measure of 1.000 when compared to itself, which indicates co-linear feature vectors, as mentioned above. Based on the magnitude of the

Table 1. Confusion Matrix for cosine similarity measure (two-tuples); accuracy 89.2%

	Actually	
Classified as	Madison	Hamilton
Madison	14	0
Hamilton	7	44

cosine similarity measures, authorship is correctly attributed to 58 of the 65 papers with a known author (here we are concerned only with those papers known to be written by either Madison or Hamilton). Of the 12 papers of disputed authorship, 11 are more similar to the Madison prototype than the Hamilton prototype. Only document number 62 is more similar to the Hamilton prototype.

In the case of three-tuples, the tuples again overlap, so each letter will be a member of three tuples. Tuple *the* is the most frequent, with a relative frequency of about 0.0009, and as before, many tuples occur with a frequency of zero. Cosine similarity results computed in the same manner as before are summarized in Table 2 and listed in detail in Table 4 in the appendix. In this case, authorship is correctly attributed to 57 of the 65 papers with a known author. Of the 12 disputed papers, all are more similar to the Madison prototype than the Hamilton prototype.

In each case these results are close to those of the classic study by Mosteller and Wallace, who found that all 12 of the disputed papers were written by Madison. At this point, we feel obligated to comment on the cost/benefit of using tuples of greater length than two. The experimental classification results of two-letter and three-letter tuples are comparable. Notice that there is little difference between the cosine similarity measures for any particular document and the two author prototypes. For example, for document 1, using two-letter tuples, the cosine similarity measure with the Madison prototype is 0.987, and with the Hamilton prototype is 0.991. Consider the delta between these values to be the absolute value of the difference between these two measures, which is, of course $|0.987 - 0.991| = 0.004$. In the case of three-letter tuples, again using document I, the similarity measures are 0.943 and 0.950, yielding a delta of 0.007. Naturally, the larger the delta value, the greater our confidence in any conclusions made based on these similarity measures. It appears that the three-letter tuples do yield better values. A box-and-whisker plot of the delta values for all pairs of similarity measures (see Fig. 1) tends to reinforce this belief, which would indicate that longer tuples are better. Consider, however, the exponential growth in the imposed computational and storage overhead. Using two-letter tuples requires $26^2 = 676$ dimensions, three-letter tuples require $26^3 = 17{,}576$ dimensions, four-letter tuples require $26^4 = 456{,}976$ dimensions, etc. As our goal was to create a powerful yet simple feature for authorship identification, we feel that this overhead significantly exceeds the benefit, and hence, we recommend the use of two-letter tuples.

3. TRANSFORMING TEXT TO IMAGES

We used the Karhunen-Loève transform to transform a feature vector into 2D coordinates, which determine a point *in* an image. This technique *is often* used in pattern rec-

Table 2. Confusion Matrix for cosine similarity measure (three-tuples); accuracy 87.6%

	Actually	
Classified as	Madison	Hamilton
Madison	14	0
Hamilton	8	43

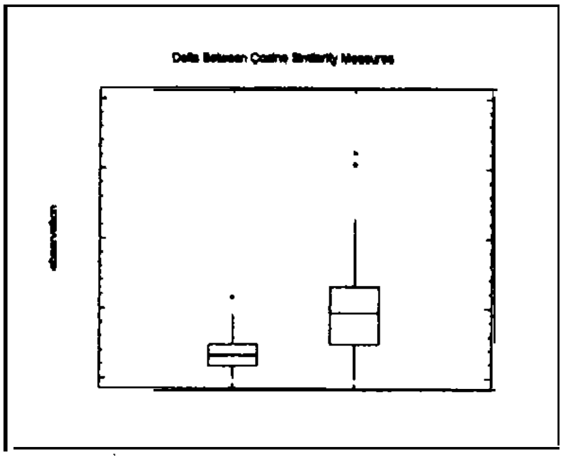

Fig. 1. Delta between cosine similarity measures.

ognition, and is explained in greater detail in textbooks on that subject (Fukunaga, 1972). The features used were the 10 two-tuples, with the greatest difference in frequency between the Hamilton and Madison prototypes. These were (in order) *er*, *to*, *ed*, *ou*, *of*, *et*, *he*, *th*, *ar*, *hi*. Only 10 features were used to ensure numerical stability in processing. Thirty-four feature vectors were computed, one for each prototype paper. The covariance matrix was computed:

$$\sum = E\{(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^t\},$$

where E is the expectation, X is a feature vector, M is the mean vector. The K-L transform was performed by finding the eigenvectors of the covariance matrix and expanding the feature vectors in terms of the eigenvectors corresponding to the largest eigenvalues. For two-dimensional representations (used here), eigenvectors corresponding to the two largest eigenvalues were used, so that for (x, y) coordinates, $x = \mathbf{X}^t\Phi_1$ and $y = \mathbf{X}^t\Phi_2$ for the two largest eigenvectors, Φ_1 and Φ_2.

The style of a text is represented as a nebula of points in a two-dimensional image, To visualize the text as a 2D image, many feature vectors for each text are created. This is done by sliding a window through a text and computing a point for each window position. Three files of nearly equal length were created out of *The Federalist Papers:* the Madison and Hamilton prototypes mentioned earlier, and all disputed papers. Each file was a large stream of text with no breaks between papers. To produce the images in Figs. 2 through 7, a 2048-character window was positioned at the beginning of a stream, then stepped through the stream in 32 character jumps. At each window position, a feature vector was calculated (the relative frequencies of the IO tuples), the feature vector was transformed into (x, y) coordinates (using the eigenvectors), and the image point at these coordinates was incremented. There were many of these points, which accumulated in the image to form a nebula. In evaluating these images for authorship determination, we use the following criteria: center of mass of the points (nebula), position of the image within the grid, and finally, shape of the image.

Figure 2 shows the image for Madison's papers; Fig. 3 shows the image for Hamilton's papers; and Fig. 4 shows the image for the unknown papers. The nebula for Madison (Fig. 2) differs from the nebula for Hamilton (Fig. 3). The Madison nebula is lower and to the right of the Hamilton nebula, and nearly the entire Madison image is located in the lower half of the grid, as compared to the Hamilton image, which is visibly shifted into the upper half of the grid. Less evident in the graphic images is that the internal structure differs: Hamilton has a central core with some diffuse outer parts: Madison has a more diffuse core and has some wispy streamers in the periphery.

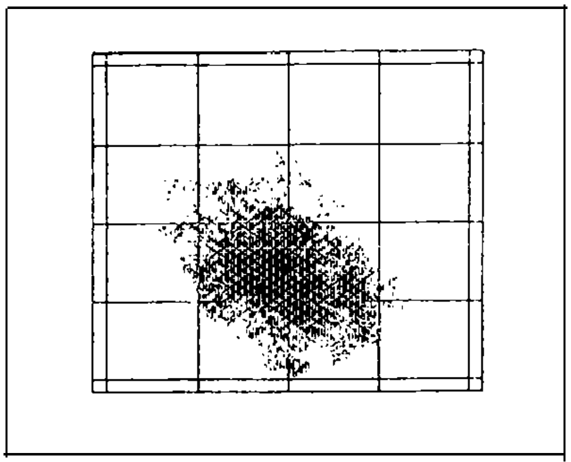

Fig. 2. Image created from Madison's papers.

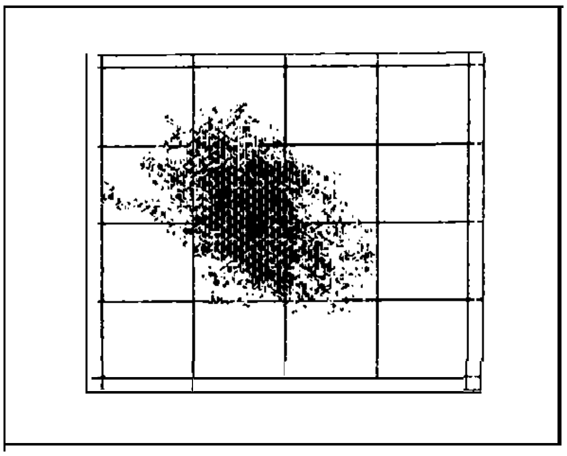

Fig. 3. Image created from Hamilton's papers

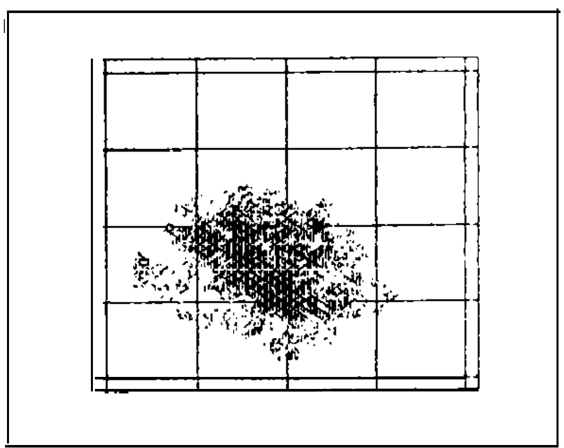

Fig. 4. Image created from the disputed papers

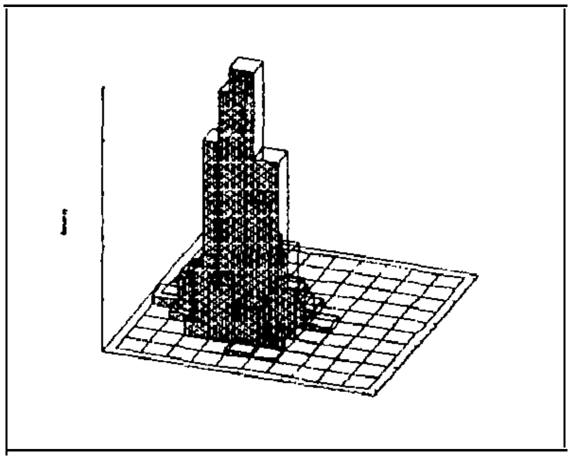

Fig. 5. Image created from Madison's papers.

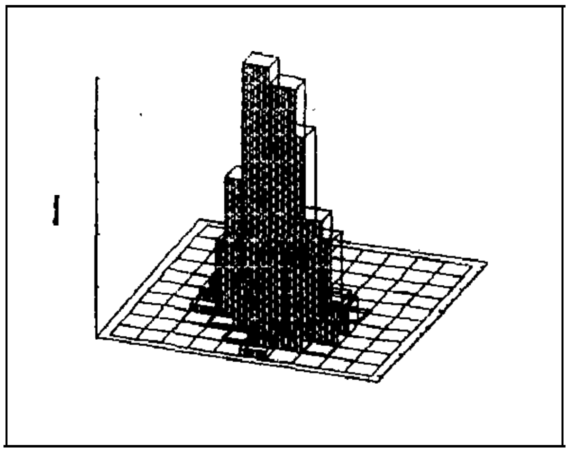

Fig. 6. Image created from Hamilton's papers.

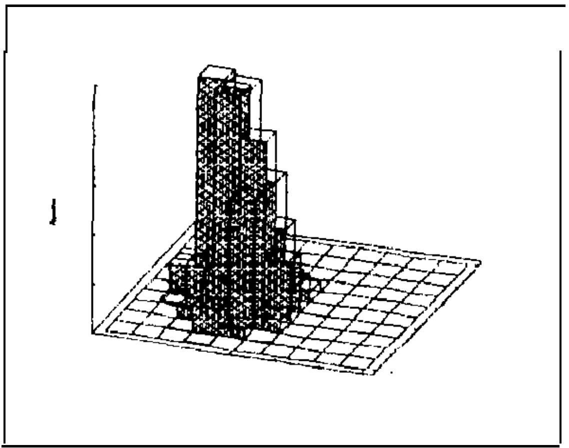

Fig. 7. Image created from the disputed papers

The third image, Fig. 4, was produced from the disputed papers. It resembles the Madison nebula more than the Hamilton nebula; the center of the nebula is at about the same location as the Madison nebula, and the core is similarly diffuse. Based on these observations, the pictures provide visual evidence that Madison wrote the unattributed papers. The same conclusion was reached with the cosine similarity measure, but the pictures provide more intuitive conclusions. In addition to the images just described, we provide three more images (Figs. 5, 6, and 7) produced from the same data, but presented as bar charts as opposed to scatter plots (scatter plots highlight the unique values, whereas bar charts provide histograms of interval values). Based on the same judgment criteria, these images tend to confirm our earlier conclusions. Additional insights into an author's style may be gained with the images. The central core of the Hamilton nebula shows an unvaried writing style; the more diffuse Madison nebula shows greater variety. It would be interesting to see if these characteristics correspond to human readers' perceptions of the authors' styles.

In the case of three-letter tuples, the 10 tuples with the greatest frequency difference between the Madison and Hamilton prototypes were (in order) *the*, *ver*, *wou*, *uld*, *oul*, *and*, *art*, *ede*, *his*, and *nce*. Images of the points generated in the same manner discussed earlier are shown in Figs. 8 through 13.

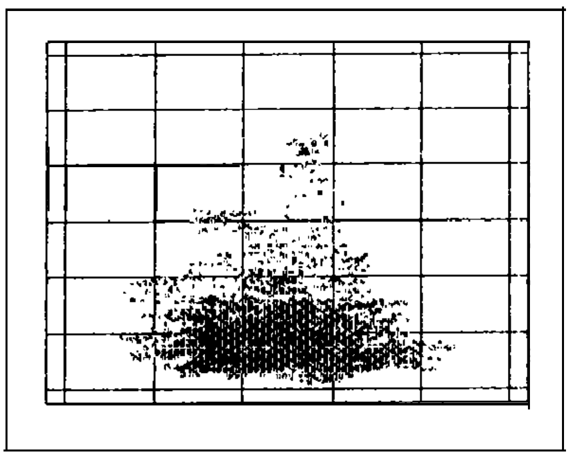

Fig. 8. Image created from Madison's papers

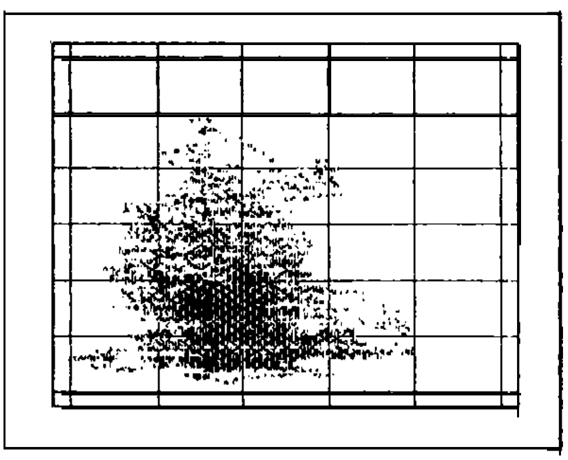

Fig. 9. Image created from Hamilton's papers

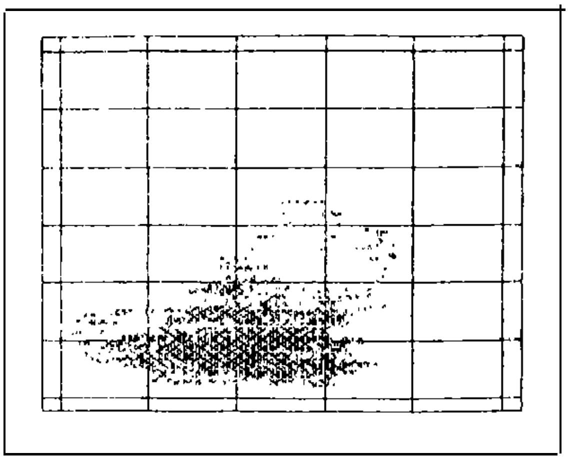

Fig. 10. Image created from the disputed papers

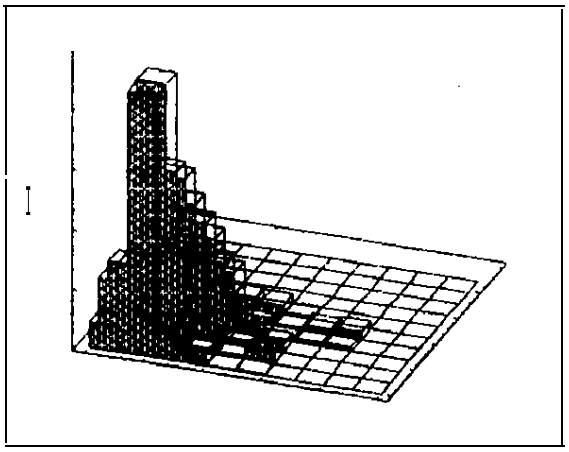

Fig. 11. Image created from Madison's papers.

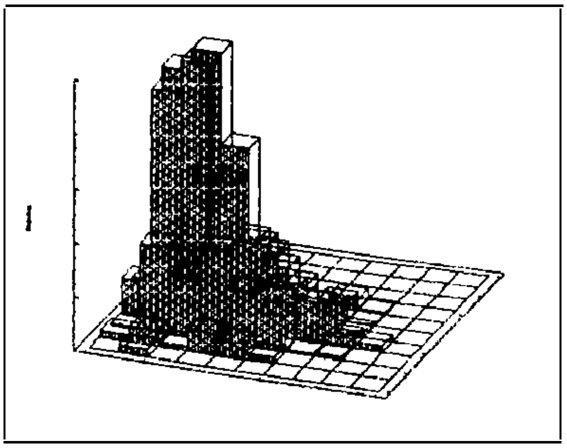

Fig. 12. Image created from Hamilton's papers.

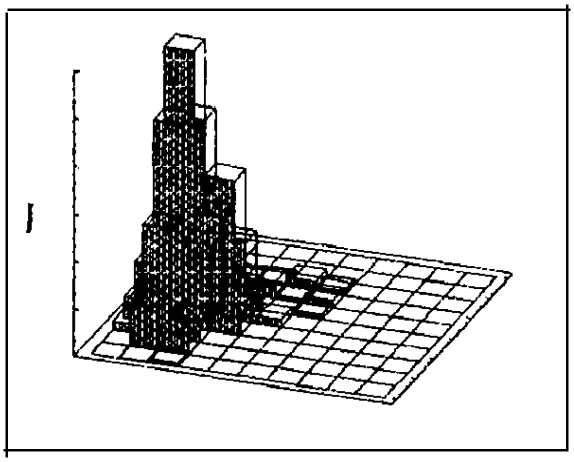

Fig. 13. Image created from the disputed papers

4. FUTURE WORK

Visualization may be used to produce images of works that are visually distinct for different authors. The method presented here uses a vast amount of information, making it unlikely that the distinction between authors is the result of a fortunate choice of features or the result of random variation. The tuple frequency method was chosen, since it provides the rich set of data necessary for generating interesting images. Other characterizations of style could be investigated and visualization techniques extended to them also. In further studies, we intend to focus on different metrics, such as tuples of greater length and different disputed document sets.

Arknowledgemenl-The authors would like to express our deep appreciation and gratitude for the assistance of the following organizations and individuals: the students and faculty of the Center for Image Analysis (CfIA), George Mason University, for their valuable comments, which contributed to the successful completion of this project. The United States Army Soldier Support Center and the United States Army Student Detachment, Fort Benjamin Harrison, Indiana, for their cooperation and assistance in reviewing this paper.

REFERENCES

Bennett, W.R., Jr. (1976). *Introduction to computer applications for non-science students.* Englewood Cliffs, NJ: Prentice-Hall.

Fukunaga, K. (1972). *Introduction to statistical pattern recognition.* New York, NY: Academic Press,

Hayes, B. (1983). Computer recreations: A progress report on the fine art of turning literature into drivel. *Scientific American*, 249. 18-28.

Hockey. S. (1980). A *guide to computer applications in the humanities.* Baltimore. MD: Johns Hopkins University Press

Holmes, D.I. (1985). The analysis of literary style—A review. *Journal of the Royal Statistical Association*, A, *148*(4), 328-341.

Kenny. A. (1986). A *stylometric study of the New Testament.* Oxford: Clarendon Press.

Kjell, B. (1985). Computational stylistics. *Sherlock Holmes: Science and literature*, pp. 20-21. Madison, WI: Division of University Outreach, University of Wisconsin-Madison.

Mendenhall, T.C. (1887). The characteristics curves of composition. Science, *9*, 237-249.

Mosteller, F., & Wallace. D.L. (1964). *Inference and disputed authorship: The Federalist.* Reading, MA: Addison. Wesley. (Also published 1984, as *Applied Bayesian and classical inference: The case of The Federalist Papers.* New York, NY: Springer-Verlag.)

Rossiter, C. (1961). *The Federalist Papers; Alexander Hamilton, James Madison, John Jay.* New York. NY: The American Library.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval.* New York, NY: McGraw-Hill.

Stratil, M.. & Oakley, R.J. (1987). A disputed authorship study of two plays attributed to Tirso de Molina. *Literary and Linguistic Computing*, 2(3), 153-160.

Tankard. J. (1986). The literary detective. *BYTE, 11*, 231-237.

APPENDIX

Table 3. Cosine similarity measure (two-tuples) between 85 papers and Madison and Hamilton prototypes

	Mad	Ham		Mad	Ham		Mad	Ham		Mad	Han,
Mad	1.000	0.995	h24	0.980	0.985	u49	0.985	0.984	h74	0.980	0.984
Ham	0.995	1.000	h25	0.988	0.991	u50	0.984	0.980	h75	0.984	0.989
h01	0.987	0.991	h26	0.987	0.992	u51	0.984	0.979	h76	0.983	0.990
j02	0.979	0.978	h27	0.985	0.988	u52	0.984	0.797	h77	0.981	0.990
j03	0.975	0.976	h28	0.985	0.987	u53	0.986	0982	h78	0.987	0.986
j04	0.970	0.974	h29	0.981	0.986	u54	0.983	0.977	h79	0.980	0.983
j05	0.961	0.966	h30	0.985	0.989	u55	0.981	0.978	h80	0.985	0.984
h06	0.984	0.988	h31	0.987	0.990	u56	0.976	0.973	h81	0.985	0.986
h07	0.987	0.990	h32	0.973	0.971	u57	0,983	0.982	h82	0.970	0.969
h08	0985	0.990	h33	0.979	0.977	u58	0.986	0.984	h83	0.985	0,987
h09	0.990	0.991	h34	0.985	0.991	h59	0.986	0.988	h84	0,990	0.989
m10	0.986	0.983	h35	0.988	0.99,	h60	0.987	0.989	h85	0.987	0.991
h11	0.978	0.986	h36	0.990	0.991	h61	0.983	0.984			
h12	0.987	0.992	m37	0.991	0.9%	u62	0.988	0.990			
h13	0.978	0.980	m38	0.994	0.992	u63	0.989	0.987			
m14	0.9%	0.989	m39	0.988	0.983	j64	0.980	0.982			
h15	0.990	0.994	m40	0.989	0.985	h65	0.983	0.988			
h16	0.984	0.989	m41	0.994	0.990	h66	0.978	0.985			
h17	0.989	0.990	m42	0.991	0.988	h67	0.977	0.980			
b18	0.979	0.975	m43	0.992	0.988	h68	0.978	0.984			
b19	0.984	0.983	m44	0.994	0.988	h69	0.988	0.989			
b20	0.980	0.978	m45	0.984	0.976	h70	0.989	0.992			
h21	0.988	0,990	m46	0.983	0.979	h71	0.984	0.988			
h22	0.991	0.994	m47	0.967	0.955	h72	0.982	0.989			
h23	0.989	0.988	m48	0.987	0.979	h73	0.984	0.987			

Table 4. Cosine similarity measure (three-tuples) between 85 papers and Madison and Hamilton prototypes

	Mad	Ham		Mad	Ham		Mad	Ham		Mad	Ham
Mad	1.000	0.986	h24	0.939	0.950	u49	0.950	0.942	h74	0.926	0.936
Ham	0.986	1.000	h25	0.951	0.959	u50	0.934	0.927	h75	0.945	0.962
h01	0,943	0.950	h26	0.952	0.961	u51	0.947	0.933	h76	0.937	0,957
j02	0.920	0.919	h27	0.941	0.947	u52	0.948	0.738	h77	0.940	0.960
j03	0.912	0.913	h28	0.949	0.95,	u53	0.948	0.939	h78	0.958	0.956
j04	0.909	0.912	h29	0.938	0.946	u54	0.939	0.932	h79	0.919	0.930
j05	0.873	0.890	h30	0.946	0.956	u55	0.942	0.937	h80	0.939	0.934
h06	0.945	0.957	h31	0.950	0.955	u56	0.911	0.908	h81	0.953	0,952
h07	0.944	0.961	h32	0.919	0.921	u57	0.941	0.936	h82	0.902	0.898
h08	0.946	0,960	h33	0.932	0.927	u58	0.946	0.941	h83	0.946	0.952
h09	0.957	0.960	h34	0.953	0.962	h59	0.949	0.950	h84	0.966	0.964
m10	0.948	0.943	h35	0.946	0.956	h60	0.954	0.964	h85	0.957	0,963
h11	0.921	0.949	h36	0.961	0.965	h61	0.940	0.946			
h12	0.949	0.964	m37	0.961	0.960	u62	0.955	0952			
h13	0.915	0.920	m38	0.975	0.965	u63	0.954	0.950			
m14	0.956	0.952	m39	0.960	0.945	j64	0.939	0.943			
h15	0.961	0.973	m40	0.968	0.952	h65	0.949	0.959			
h16	0.949	0.960	m41	0.973	0.963	h66	0.945	0.955			
h17	0.956	0.960	m42	0.958	0.948	h67	0.932	0.936			
b18	0.929	0.920	m43	0.973	0.959	h68	0.931	0.942			
b19	0.937	0,934	m44	0.970	0.953	h69	0.950	0.954			
b20	0.930	0.923	m45	0.959	0.938	h70	0.957	0.967			
h21	0.953	0.959	m46	0.951	0.939	h71	0.948	0.957			
h22	0.967	0.976	m47	0.877	0.848	h72	0.938	0.959			
h23	0.952	0.950	m48	0.945	0.925	h73	0.942	0.955			