

# Automated Scoring of Handwritten Essays Based on Latent Semantic Analysis

Sargur Srihari, Jim Collins, Rohini Srihari,  
Pavithra Babu, and Harish Srinivasan

Center of Excellence for Document Analysis and Recognition (CEDAR)  
University at Buffalo, State University of New York Amherst, New York 14228, U.S.A  
[srihari@cedar.buffalo.edu](mailto:srihari@cedar.buffalo.edu)

**Abstract.** Handwritten essays are widely used in educational assessments, particularly in classroom instruction. This paper concerns the design of an automated system for performing the task of taking as input scanned images of handwritten student essays in reading comprehension tests and to produce as output scores for the answers which are analogous to those provided by human scorers. The system is based on integrating the two technologies of optical handwriting recognition (OHR) and automated essay scoring (AES). The OHR system performs several pre-processing steps such as forms removal, rule-line removal and segmentation of text lines and words. The final recognition step, which is tuned to the task of reading comprehension evaluation in a primary education setting, is performed using a lexicon derived from the passage to be read. The AES system is based on the approach of latent semantic analysis where a set of human-scored answers are used to determine scoring system parameters using a machine learning approach. System performance is compared to scoring done by human raters. Testing on a small set of handwritten answers indicate that system performance is comparable to that of automatic scoring based on manual transcription.

## 1 Introduction

Handwritten essays are widely used for student performance evaluation in schools and colleges. Since this approach to evaluation is efficient and reliable it is likely to remain a key component of learning. Assessing large numbers of handwritten essays is a relatively time-consuming and monotonous task. At the same time there is an intense need to speed up and enhance the process of rating handwritten essays while maintaining cost effectiveness. The assessment can also be used as a source of timely, relatively inexpensive and responsible feedback about writing. The paper describes a first attempt at designing a system for reading, scoring and analyzing handwritten essays from large scale assessments to provide assessment results and feedback. Success in designing such a system will not only allow instructors to provide timely feedback to students but also can provide feedback to education researchers and educators.

Writing done by hand is the primary means of testing students on state assessments. Consider as an example the New York State English Language Assessment (ELA) administered statewide in grades 4 and 8. In the reading part of the test the student is asked to read a passage such as that given in Fig 1 and answer several questions in writing.

**American First Ladies**

When George Washington was elected the first President of the United States, there was much discussion about the role his wife, Martha, should play. "Lady Washington," the people began to call her. But Martha did not want to be treated like royalty or aristocracy. She decided that she would be an equal partner with her husband on social occasions. By doing this, she established a very important role for the President's wife—"hostess for the nation."

That role was expanded by James Madison's wife, Dolley Madison, the first woman to be called "First Lady." During the early 1800s, women had few rights. They could not vote or take part in politics. Married women were not even allowed to own property or make a will. A woman's education was usually limited to homemaking skills. But Dolley Madison came from a Quaker family whose community opened its schools to boys and girls. Dolley became an outgoing woman with strong opinions, whose influence on her husband was well known. She was also considered to be the center of society in Washington, D.C., hosting receptions at which politicians and diplomats gathered along with the general public. After her death in 1849, Dolley was honored by President Zachary Taylor, who called her "my only First Lady for half a century."

Many other First Ladies had strong influences on their husbands, but a woman who really made the job her own was Anna Eleanor Roosevelt, the wife of Franklin Delano Roosevelt. Just two days after her husband's inauguration in 1933, Eleanor held the first press conference ever given by a presidential wife. During FDR's presidency, Eleanor was always there with suggestions, proposals, and ideas. Sharecroppers' parent workers, students, and other people whom she had encountered on her travels were brought to Washington to meet the President.

\*sharecropper: tenant farmer

---

**Some Famous First Ladies in History**

Timeline of historical events:

- 1776: Declaration of Independence signed
- 1789: Constitution becomes law
- 1800: First Presidential Inauguration
- 1803: Louisiana Purchase
- 1820: Missouri Compromise
- 1849: Dolley Madison's death
- 1861: Civil War begins
- 1869: 15th Amendment passed
- 1870: 13th Amendment passed
- 1876: Reconstruction ends
- 1888: 15th Amendment passed
- 1890: 19th Amendment passed
- 1900: 16th Amendment passed
- 1913: 17th Amendment passed
- 1920: 19th Amendment passed

Portraits of: Martha Washington (1775-1802), Eleanor Roosevelt (1884-1962), Sarah Wilson (1812-1841), Eleanor Roosevelt (1884-1962).

Page 12 | Session 2: Part 1 | Page 13

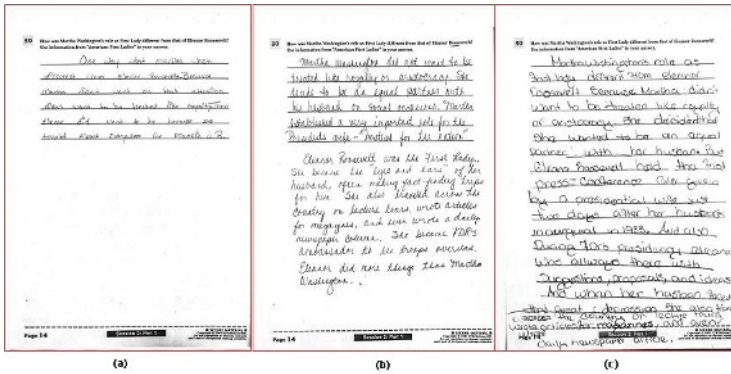
Fig. 1. From the New York English Language Arts assessment for Grade 8, 2001 – two of three pages of the story “American First Ladies” are shown

An example of a reading comprehension question based on the passage of Fig. 1 is the following: “How was Martha Washington’s role as First Lady different from that of Eleanor Roosevelt? Use information from American First Ladies in your answer.” The completed answer sheets of three different students to the question are given in Fig. 2. The answers are scored by human assessors on a seven-point scale of 0-6.

There is significant practical and pedagogical value in computer-assisted evaluation of such tests. The task of scoring and reporting the results of these assessments in a timely manner is difficult and relatively expensive. There is also an intense need to test later in the year for the purpose of capturing the most student growth and at the same time meet the requirement to report student scores before summer break. The biggest challenge is that of reading and scoring the handwritten portions of large-scale assessments.

From the research viewpoint an automated solution will allow studying patterns among handwritten essays that may be otherwise laborious or impossible. For instance metrics can be obtained for identifying difficulties struggling students are having, for measuring repetition of sections from the original passage, for identifying language constructs specific to the population, etc.

The assessment problem is a well-constrained problem in artificial intelligence whose solution will push forward existing technologies of handwriting recognition and automatic essay scoring. Much of artificial intelligence research has



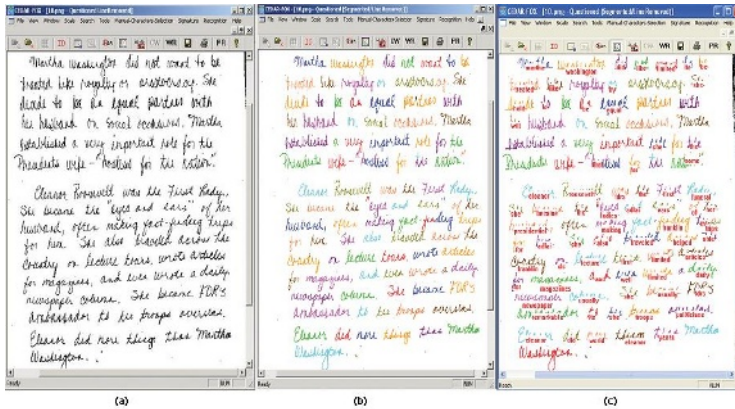
**Fig. 2.** Sample answer sheets of three students (a-c) based on the reading comprehension passage of Fig. 1. The human assigned scores for these essays, on a scale of 0-6, were 2, 4 and 4 respectively.

progressed in the quest for solutions for specific problems, and this problem promises to be an exciting one both in terms of the task and its use. Solving the problem also promises to reduce costs and raise efficiency of large-scale assessments, which use handwritten essays.

## 2 Component Technologies and Previous Work

The first step is that of computer reading of handwritten material in the scanned image of an answer booklet page, or optical handwriting recognition (OHR). While computers have become indispensable tools for two of three R's, viz., arithmetic and writing, their use in the third R of reading is still emerging. OHR involves several processing steps such as form (or rule line) removal, line/word segmentation and recognition of individual words. Word recognition relies on a lexicon of words—which could be derived from the passage, question and rubric available in statewide tests. The result of processing the scanned handwritten answer of Fig.2(a) by an OHR system is shown in Fig. 3. The lexicon consisted of words from the passage: 1800s, 1849, 1921, 1933, 1945, 1962, 38000, a, able, about, across, adlai, after, allowed, along, also, always, ambassador, American, etc.

When the lexicon is limited, a majority of the words are correctly recognized although there are errors in some recognized words, and some words are entirely missed. These errors can be reduced by better word segmentation and by using linguistic context in the form of transitional probabilities between words, between parts-of-speech tags, etc. It is possible that certain words, when considered in isolation, are illegible. Local context can resolve such ambiguity. The reading comprehension passage and the rubric provide a rich source of contextual information that can be exploited to get high recognition rates. However, the task itself is one of correct interpretation rather than that of recognizing every illegible word.



**Fig. 3.** Results of processing the answer sheet shown in Fig. 2(b) by the OHR system: (a) after forms and rule-line removal, (b) after line and word segmentation (words are shown in different colors), and (c) after word recognition—the word recognizer used a lexicon from the rubric of the question for performing word recognition; only high confidence results are shown

Two basic approaches to automatic essay scoring (AES) are: IR (Information Retrieval)- based and linguistics-based. The IR approach is exemplified by latent semantic analysis (LSA) where keywords and their co-occurrence statistics are used to uncover hidden semantic links between a gold standard and the essay to be evaluated. Linguistic approaches emphasize the use of structures to decode the semantics.

Each of the two areas of OHR and AES has a significant research literature, with virtually no overlap. This research involves integrating concepts and algorithms from each of these areas, which are briefly summarized below.

## 2.1 Optical Handwriting Recognition (OHR)

OHR is concerned with transforming an image of handwritten text into its textual form. It includes character recognition (OCR), word recognition, part-of-speech tagging, and style recognition. A survey of both on-line (also called dynamic) and off-line (or static or optical) handwriting recognition is [10]. While dynamic handwriting recognition has enjoyed considerable success due to the increasing popularity of tablet PCs and PDAs, OHR has enjoyed success in constrained domains such as postal addresses. To distinguish the two types of handwriting recognition the off-line case is also referred to as optical handwriting recognition (OHR). The higher complexity of OHR stems from the lack of temporal information and the complexity of document analysis.

Prior to OHR there are several pre-processing steps to be performed on document images, e.g., detecting and eliminating extraneous information such as logos, ruled lines and margin lines. Within the handwritten text the ordering of the lines has to be determined and within each line the words need to be segmented.

A system for reading unconstrained handwritten pages known as PENMAN was developed by us [13] which has since been developed into the CEDAR-FOX system for the analysis of handwritten documents for forensic analysis [14] and the CEDAR-OHR system. These systems have tools for gray-scale to binary thresholding, rule line removal, and line/word segmentation. There are interactive user interfaces available for the analysis of the documents by researchers.

*Character and Word Recognition:* Recognition of characters and words is performed in a two step process of feature extraction followed by classification. Features can be either the raw image pixels or shape descriptors. Features can be at the character level (called analytic recognition) or at the word level (holistic recognition). Word recognition can be performed reasonably well for correctly segmented words with small lexicons. The process is error prone for mis-segmented text and large lexicons. Exploiting statistical dependencies between words was first explored in the OCR domain [3] and then extended to on-line handwriting [12]. There has been considerable success in constrained-domain handwriting recognition, e.g., in handwritten address interpretation [7]. Statistical dependencies between word tags corresponding to parts of speech (POS) rather than to words themselves was also explored by our group. In fact the proposed research will expand on this idea but focus on NEs rather than POS.

*Handwriting Interpretation:* Handwriting Interpretation is a goal-oriented task where the goal is not so much one of recognizing every character and word perfectly but to perform the overall task in an accurate manner. It involves using basic handwriting recognition tools together with contextual information to solve specific tasks even when there is significant uncertainty in the specific components. Such approaches have found success when the domain is limited and contextual information is available. For instance, in the domain of postal addresses a system was developed for determining the destination irrespective of whether the individual components were correctly written [15]. The strategy was to recognize the most easily recognizable parts of the address first, which in this case consists of the ZIP Code and street number. These two “islands” are used to narrow down the lexicon of choices of the street name, which simplifies the task of recognizing the street name. The recognition of the ZIP Code is constrained by the state name and city name. The mutual constraints lead to a correct interpretation irrespective of spelling errors, mistakes and illegibility. Today, over 90% of all handwritten addresses in the United States are interpreted by OHR. This triangulation is useful for recognition of essay words when constraints imposed by certain words can be used to disambiguate illegible words.

## 2.2 Automatic Essay Scoring (AES)

Automated essay scoring has been a topic of research over four decades. A limitation of all past work is that the essays or examinations have to be in computer readable form. A survey of previous AES methods has been made by Palmer, et. al (2002). Project Essay Grade (PEG) (Page, 1961) uses linguistic features from which a multiple regression equation is developed. In the Production Automated

Essay Grading System a grammar checker, a program to identify words and sentences, software dictionary, a part-of-speech tagger, and a parser are used to gather data. E-rater (Burstein, 2003) uses a combination of statistical and NLP techniques to extract linguistic features. Larkey (1998) implemented an AES approach based on text categorization techniques (TCT).

A promising approach to AES is based on a technique developed in the information retrieval community known as latent semantic indexing. Its application to AES, known as latent semantic analysis (LSA), uncovers lexical semantic links between an essay and a gold standard. Landauer, et. al. (1998) developed the Intelligent Essay Assessor using LSA. A matrix for the essay is built, and then transformed by the algebraic method of singular value decomposition (SVD) to approximately reproduce the matrix using reduced dimensional matrices built for the topic domain. Using SVD new relationships between words and documents are uncovered, and existing relationships are modified to represent their significance. Using LSA the similarity between two essays can be measured despite differences in individual lexical items. Where a gold standard exists, LSA is a robust approach. It correlates as closely with human raters as human raters correlate with each other (Landauer et al. 2003).

Hybrid systems combine LSA word vector similarity metrics with structure-based linguistic features. One such system is the widely-used ETS E-rater (Burstein, 2003) which has been used to grade essays on the GMAT, TOEFL, and GRE exams. E-rater is trained on about 300 hand-graded essays for each question or prompt. A predictive statistical model is developed from the set of scored essays by beginning with a set of 50-70 features and using stepwise linear regression to select features necessary to assign a score from 1-6. E-rater uses part of speech (POS) tagging and shallow parsing to identify subjunctive auxiliary verbs, more complex clause types (e.g. complement, infinitive, and subordinate clauses), and discourse cue words (e.g. because, in summary, for example). Discourse cue words are used both as individual features and as a way to divide essays into labeled discourse segments. The score is a combination of an overall score and the scores for each discourse segment. Additional features include the presence of words like possibly and perhaps, presence of sentence-initial infinitive phrases, and discourse-deictic words like this and these. To evaluate content and word choice, E-rater uses vectors of word frequencies. The training set is collapsed into six categories (one for each scoring value), and each discourse segment is scored by comparing it with the six categories. The mean of the argument scores is adjusted for the number of arguments (to penalize shorter essays).

### 3 System Architecture

The basic software structure consists of OHR and AES systems. They are integrated into a single automated assessment platform, which can both assign scores and provide analytic feedback. The system is designed to be interactive so that the user— a researcher or educator— will be able to examine the decisions made, and accept or change the decisions.

The OHR system has two phases/components: (i) Rubric Processor: a system that extracts lexicons, entities and concepts from the reading passage, question and rubric– it compiles various indices including lexicons of words that can be used as features by the answer OHR, and (ii) Answer Processor: a system to analyze and score the scanned answer sheet(s) The answer sheets are scanned as gray scale images at a resolution of 300 pixels per inch. The OHR system is written in C with visual interfaces in Visual C++.

The AES system also has two phases. In the training phase the system parameters are learnt from a set of human-scored samples. In the testing phase these parameters are used in performing the evaluation.

### 3.1 OHR

Recognition of the scanned document begins with a number of document image pre-processing steps such as extracting the foreground from the background, eliminating non-informative material such as rule lines and other printed markings, determining the presence of handwritten words, their reading sequence, etc. Examples of line/word segmentation and recognition, as performed by the CEDAR OHR system, were seen earlier in Fig. 3. The main tasks for the OHR system are as follows:

(1) Word segmentation: The tasks here are the segmentation of lines and words of handwritten text in the presence of ambiguity. To determine whether a gap is a true gap or not features are taken into account from the current document rather than solely rely on a learning set.

(2) Word recognition: When vocabularies provided to the word recognizer are large contextual information needs to be exploited to dynamically limit word choices during the process of recognition. Contextual information is readily available to the recognizer in the form of the passage to be read and the answer rubric.

After the words in the scanned answer documents are recognized by the OHR system the resulting word sequences are written to text files. These text files are then pre-processed for AES which include the following steps.

- a. Removing punctuations and special characters
- b. Converting upper case to lower case for generalization
- c. Stop word removal - removing common words such as *a* and *the* which occur very often and are not of significant importance
- d. Stemming - morphological variants of words have similar semantic interpretations and therefore a stemming algorithm is used to reduce the word to its stem or root form. The algorithm [11] uses a technique called suffix stripping where an explicit suffix list is provided along with a condition on which the suffix should be removed or replaced to form the stem of the word, which would be common among all variations. For example the word *reading* after suffix stripping is reduced to *read*.

### 3.2 AES

When a large number of student answers are available, which is the case in statewide assessments, a holistic approach will assist in providing computer automated grades close to those that would be assigned by expert human graders. The Latent Semantic Analysis (LSA) approach can take advantage of the availability of a large training corpus. The training corpus consists of human-scored answer documents. The potentiality of LSA to imitate human judgment has been explored and found to have a strong correlation [4].

The underlying semantics of the training corpus are extracted using LSA and without the use of any other external knowledge. The method captures how the variations in term choices and variations in answer document meanings are related. However, it does not take into consideration the order of occurrence of words. This implies that even if two students have used different words to convey the same message, LSA can capture the co-relation between the two documents. This is because LSA depicts the meaning of a word as an average of the connotation of the documents in which it occurs. It can similarly judge the correctness of an answer document as an average of the measure of correctness of all the words it contains.

Mathematically this can be explained as the simultaneous representation of all the answer documents in the training corpus as points in semantic space, with initial dimensionality of the order of the number of terms in the document. This dimensionality is reduced to an optimal value large enough to represent the structure of the answer documents and small enough to facilitate elimination of irrelevant representations. The answer document to be graded is also placed in the reduced dimensionality semantic space and the by and large term-based similarity between this document and each of those in the training corpus can then be determined by measuring the cosine of the angle between the two documents at the origin.

A good approximation of the computer score to a human score heavily depends on the optimal reduced dimensionality. This optimal dimension is related to the features that determine the term meaning from which we can derive the hidden correlations between terms and answer documents. However a general method to determine this optimal dimension is still an open research problem. Currently a brute force approach is adopted. Reducing the dimensions is done by omitting inconsequential relations and retaining only significant ones. A factor analysis method such as Singular Value Decomposition (SVD) helps reduce the dimensionality to a desired approximation.

The first step in LSA is to construct a  $t \times n$  term-by-document matrix  $M$  whose entries are frequencies. SVD or two-mode factor analysis decomposes this rectangular matrix into three matrices [1]. The SVD for a rectangular matrix  $M$  can be defined as

$$M = TSD', \quad (1)$$

where prime (') indicates matrix transposition,  $M$  is the rectangular term by document matrix with  $t$  rows and  $n$  columns,  $T$  is the  $t \times m$  matrix, which



**Table 1.** An Example 154 x 31 Term by Document Matrix  $M$ , where  $M_{ij}$  is the frequency of the  $i^{th}$  term in the  $j^{th}$  answer document

Term/Doc	D1	D2	D3	D4	D5	D6	D7	D8	...	D31
T1	0	0	0	0	0	0	0	0	...	0
T2	2	1	2	1	3	1	2	1	...	2
T3	0	1	0	1	3	1	2	1	...	1
T4	0	1	0	0	0	1	0	0	...	0
T5	0	1	0	0	0	0	1	0	...	0
T6	0	1	1	0	0	1	2	0	...	0
T7	0	0	0	0	1	0	0	0	...	0
T8	0	0	0	0	1	0	1	0	...	0
T9	...	...	...	...	...	...	...	...	...	...
T154	0	0	0	0	0	0	0	0	...	0

describes rows in the matrix  $M$  as left singular vectors of derived orthogonal factor values,  $D$  is the  $n \times m$  matrix, which describes columns in the matrix  $M$  as right singular vectors of derived orthogonal factor values,  $S$  is the  $m \times m$  diagonal matrix of singular values such that when  $T$ ,  $S$  and  $D'$  are matrix multiplied  $M$  is reconstructed, and  $m$  is the rank of  $M = \min(t, n)$ .

To reduce the dimensionality to a value, say  $k$ , from the matrix  $S$  we have to delete  $m - k$  rows and columns starting from those which contain the smallest singular value to form the matrix  $S_1$ . The corresponding columns in  $T$  and rows in  $D'$  are also deleted to form matrices  $T_1$  and  $D'_1$  respectively. The matrix  $M_1$  is an approximation of matrix  $M$  with reduced dimensions as follows

$$M_1 = T_1 S_1 D'_1. \quad (2)$$

Standard algorithms are available to perform SVD. To illustrate, a document-term matrix constructed from 31 essays from the *American First Ladies* example shown in Fig 1 and Fig 2 are given in Table 1. Since the corpus contains 31 documents with 154 unique words,  $M$  has dimensions  $t = 154$  and  $m = 31$ .

**Training Phase.** The following steps are performed in the training phase:

- 1) Answer documents are preprocessed and tokenized into a list of words or terms— using the document pre-processing steps described in section 3.1.
- 2) An *Answer Dictionary* is created which assigns a unique file ID to all the answer documents in the corpus.
- 3) A *Word Dictionary* is created which assigns a unique word ID to all the words in the corpus.
- 4) An *index* with the word ID and the number of times it occurs (word frequency) in each of the 31 documents is created.
- 5) A *Term-by-Document Matrix*,  $M$  is created from the index, where  $M_{ij}$  is the frequency of the  $i$ th term in the  $j$ th answer document.

**Validation Phase.** A set of human graded documents, known as the validation set, are used to determine the optimal value of  $k$ . Each of them are passed as

query vectors and compared with the training corpus documents. The following steps are repeated for each document.

- 1) A vector  $Q$  of term frequencies in the query document is created, similar to the way  $M$  was created
- 2)  $Q$  is then added as the 0th column of the Matrix  $M$  to give a matrix  $M_q$
- 3) SVD is performed on the matrix  $M_q$ , to give the  $TSD'$  matrices
- 4) Steps 5-10 are repeated for dimension values,  $k = 1$  to  $\min(t, m)$
- 5) Delete  $m - k$  rows and columns from the  $S$  matrix, starting from the smallest singular value to form the matrix  $S_1$ . The corresponding columns in  $T$  and rows in  $D'$  are also deleted to form matrices  $T_1$  and  $D'_1$  respectively
- 6) Construct the matrix  $M_{q1}$  by multiplying the matrices  $T_1 S_1 D'_1$
- 7) The similarity between the query document  $x$  (the 0th column of the matrix  $M_{q1}$ ) and each of the other documents  $y$  in the training corpus (subsequent columns in the matrix  $M_{q1}$ ) are determined by the cosine similarity measure defined as

$$\text{CosineSimilarity} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i} \sqrt{\sum_{i=1}^n y_i}} \quad (3)$$

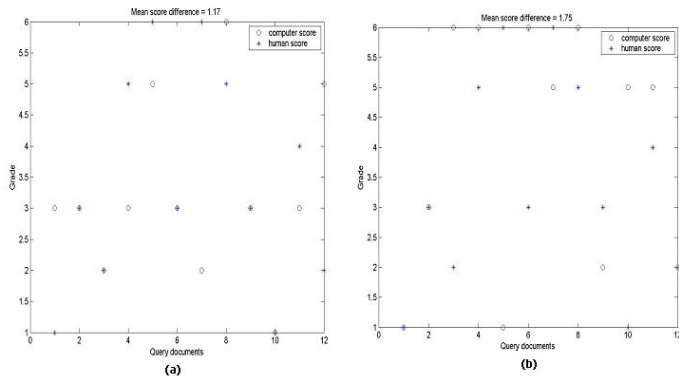
- 8) The training documents with the highest similarity score, when compared with the query answer documents are selected and the human scores associated with these documents are assigned to the documents in question respectively
- 9) The mean difference between the LSA graded scores and that assigned to the query by a human grader is calculated for each dimension over all the queries
- 10) Return to step 4
- 11) The dimension with least mean difference is selected as the optimal dimension  $k$  which is used in the testing phase.

**Testing Phase.** The testing set consists of a set of scored essays not used in the training and validation phases. The term-document matrix constructed in the training phase and the value of  $k$  determined from the validation phase are used to determine the scores of the test set.

## 4 Experimental Results

The corpus for experimentation consisted of 71 handwritten answer essays. Of these essays 48 were by students and 23 were by teachers. Each of the 71 answer essays were manually assigned a score by education researchers. The essays were divided into 47 training samples, 12 validation samples and 12 testing samples. The training set had a human-score distribution on the seven-point scale as follows: 1,8,9,10,2,9,8. Both the validation and testing sets had human-score distributions of 0,2,2,3,1,2,2.

Two different sets of 71 transcribed essays were created, the first by manual transcription (MT) and the second by the OHR system. The lexicon for the OHR system consisted of unique words from the passage to be read, which had a size



**Fig. 4.** Comparison of human scores, MT-LSA scores and OHR-LSA scores on 12 student responses to the *American First Ladies* question: (a) MT-LSA scores (open circles) are within 1.17 of human scores (stars), and (b) OHR-LSA scores (open circles) are within 1.75 of human scores (stars)

of 274. Separate training and validation phases were conducted for the MT and OHR essays. For the MT essays, the document-term matrix  $M$  had  $t = 490$  and  $m = 47$  and the optimal value of  $k$  was determined to be 5. For the OHR essays, the corresponding values were  $t = 154$ ,  $m = 47$  and  $k = 8$ . The smaller number of terms in the OHR case is explained by the fact that several words were not recognized.

Comparisons of the human-assigned scores (the gold-standard) with (i) automatically assigned scores based on MT is shown in Fig. 4(a) and (ii) automatically assigned scores based on OHR is shown in 4(b). In both plots the human scores are shown as stars (\*) and machine-assigned scores as open circles. Using MT the human-machine mean difference was 1.17 (Fig. 4 (a)). This is consistent with a one-point difference between LSA and the gold-standard previously established in large scale testing. Using OHR the the human-machine difference was 1.75 (Fig. 4 (b)). Thus a 0.58 difference is observed between MT and OHR using LSA scoring. Although the testing set is small, these preliminary results demonstrate the potential of the method for holistic scoring and robustness with OHR errors.

## 5 Summary and Discussion

An approach to automatically evaluating handwritten essays in reading comprehension tests has been described. The design is based on optical handwriting recognition (OHR) and automatic essay scoring (AES). The lexicon for OHR is obtained from the passage to be read by the students. The AES method is based on latent semantic analysis (LSA) which is a holistic method of scoring that has shown much promise. Results on a small testing set show that with manually transcribed (MT) essays, LSA scoring has about a one-point difference from hu-

man scoring which is consistent with large scale testing. With the same test set, OHR-LSA scoring has a half-point difference from MT-LSA scoring.

The results point out that despite errors in word recognition the overall scoring performance is good enough to have practical value. This points out that when the evaluation of an OHR system is based not so much on word recognition rates but in terms of the overall application in which it is used, the performance can be quite acceptable. The same phenomenon has been observed in other OHR applications such as postal address reading where the goal is not so much as to read every word correctly but achieve a correct sortation.

The LSA approach has some disadvantages for its use in a class-room setting since it is based on a holistic approach. Like other IR approaches based on a “bag of words” model, LSA ignores linguistic structures. While the scoring is based on content, it is not based on idea development, organization, cohesion, style, grammar, or usage conventions. Such an approach, known as analytic scoring, will need to take into account linguistic structures. The result of analytic scoring will be more useful to teachers and education researchers. Future work will involve improving the robustness of OHR, particularly in segmenting touching text lines, and the use of information extraction techniques for tighter integration between OHR and AES.

## References

1. Baeza-Yates, R. and Ribeiro-Neto, B.: Modern information retrieval. New York: Addison-Wesley (1999).
2. Burstein, J.: The E-rater Scoring Engine: Automated essay scoring with natural language processing. In Automated Essay Scoring (2003).
3. J. J. Hull: “Incorporation of a Markov model of syntax in a text recognition algorithm,” in Proceedings of the Symposium on Document Analysis and Information Retrieval, pp. 174-183 (1992).
4. Landauer, T. and D. Laham and P. Foltz.: Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Automated Essay Scoring (2003).
5. Landauer, T.K., P. W. Foltz and D. Laham: An introduction to latent semantic analysis, *Discourse Processes*, 25, pp. 259-284.
6. Larkey, L.S.: “Automatic essay grading using text categorization techniques,” Proceedings ACM-SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 90-95.
7. Mahadevan, U., and Srihari, S.N.: “Parsing and recognition of city, state and ZIP Codes in handwritten addresses,” in Proceedings of Fifth International Conference on Document Analysis and Recognition (ICDAR), Bangalore, India, pp. 325-328.(1999)
8. Page, E. B.: “Computer grading of student prose using modern concepts and software,” *Journal of Experimental Education*, 62, pp. 127-142.
9. Palmer, J., R. Williams and H. Dreher: “Automated essay grading system applied to a first year university subject - how can we do better?” *Informing Science*, pp. 1221-1229 June 2002.
10. Plamondon, R., and S. N. Srihari.: “On-line and off-line handwriting recognition: A comprehensive survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1): 63-84, 2000.

11. Porter, M.F.: "An Algorithm for Suffix Stripping" Program, 14(3),pp. 130-137, 1980.
12. Srihari, R. K., S. Ng, C.M. Baltus and J. Kud: "Use of language models in on-line sentence/phrase recognition," in Proceedings of the International Workshop on Frontiers in Handwriting Recognition, Buffalo, pp. 284-294, 1993.
13. Srihari, S. N., and Kim, G.: "PENMAN: A system for reading unconstrained handwritten page images," in Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 97), Annapolis, MD, pp. 142-153, 1997.
14. Srihari, S. N., B. Zhang, C. Tomai, S. Lee, Z. Shi and Y. C. Shin: "A system for handwriting matching and recognition, " in Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 03), Greenbelt, MD, pp. 67-75, 2003.
15. Srihari, S.N., and E. J. Keubert: "Integration of handwritten address interpretation technology into the United States Postal Service Remote Computer Reader System," Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR 97), Ulm, Germany, pp. 892-896, 1997.