Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

Enabling data mining of handwritten coursework

Thomas F. Stahovich [a], Hanlung Lin [b,1]

[a] Department of Mechanical Engineering, University of California, Riverside, CA 92521, USA
[b] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

A R T I C L E I N F O

Article history:
Received 30 September 2015
Received in revised form
15 January 2016
Accepted 18 January 2016

Keywords:
Single-stroke classification
Stroke grouping
Educational data mining
Sketch understanding

A B S T R A C T

Data mining has become an increasingly important tool for education researchers and practitioners. However, work in this field has focused on data from online educational systems. Here, we present techniques to enable data mining of handwritten coursework, which is an essential component of instruction in many disciplines. Our techniques include methods for classifying pen strokes as diagram, equation, and cross-out strokes. The latter are used to strike out erroneous work. We have also created techniques for grouping equation strokes into equation groups and then individual characters. Our results demonstrate that our classification and grouping techniques are more accurate than prior techniques for this task. We also demonstrate applications of our techniques for automated assessment of student competence. We present a novel approach for measuring the correctness of exam solutions from an analysis of lexical features of handwritten equations. This analysis demonstrates, for example, that the number of equation groups correlates positively with grade. We also use our techniques to extend graphical protocol analysis to free-form, handwritten problem solutions. While prior work in a laboratory setting suggests that long pauses are indicative of low competence, our work shows that the frequency of long pauses during exams correlates positively with competence.

1. Introduction

Data mining has become an increasingly important tool for education researchers and practitioners [39]. For example, data mining has been used to identify patterns of learning behavior [25] and for assessing student learning [13]. Progress in this field has been accelerated by the ready availability of data from online educational systems such as intelligent tutoring systems [49] and learning management systems like Moodle and Blackboard [26].

However, in many disciplines such as engineering, science, and math, paper-and-pencil problem-solving is still a fundamental component of instruction. Our goal is to develop approaches for mining handwritten coursework. One of the primarily challenges in accomplishing this is extracting semantic information from digital ink data. Online systems have been a convenient target of data mining because they typically produce log files with structures that facilitate parsing. By contrast, extracting semantic information from digital ink data is a much more challenging problem as it requires recognition rather than parsing.

Here we present techniques to enable data mining of students' handwritten solutions to engineering problems. These solutions typically include both diagrams and equations. Diagrams provide an abstraction of the problem that guides the construction of the equations. For our work, we consider a database of solutions to problems from an undergraduate Statics course. Statics is the subdiscipline of engineering mechanics focused on the equilibrium of structures subjected to forces. Our database comprises nearly 60,000 pages of ink with 28×10^6 time-stamped pen strokes from 700 students. This data was recorded with Livescribe smartpens.

Fig. 1 shows a typical Statics problem and Fig. 2 shows the sort of solution a student might produce. The solution includes a free body diagram and equilibrium equations. The former represents the forces acting on the system under consideration, while the latter are the application of Newton's second law. Fig. 2 also shows examples of cross-out marks used to strike out erroneous work.

Our techniques perform two primary tasks. They first classify the strokes into one of three categories: equation, free body diagram, and cross-out. Then they cluster the equation pen strokes, first into equation groups and then into individual characters. An equation group is a string of characters belonging to a single equation and written on the same baseline. The pen strokes in Fig. 2 are color-coded to differentiate the three kinds of pen strokes. Fig. 9 shows examples of equation groups.

Our techniques are intended to support applications in educational data mining. For example, Van Arsdale and Stahovich [52] developed methods that predict the correctness of handwritten solutions to Statics problems from the spatial and temporal organization of the work. They characterize a solution with various

[1] Current address: Amazon, 207 Boren Ave N, Seattle, WA 98109, USA.

http://dx.doi.org/10.1016/j.cag.2016.01.002
0097-8493/© 2016 Published by Elsevier Ltd.

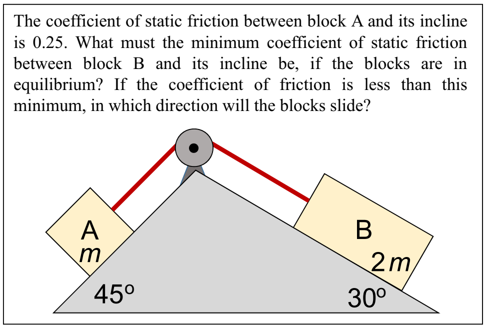

Fig. 1. A typical Statics problem.

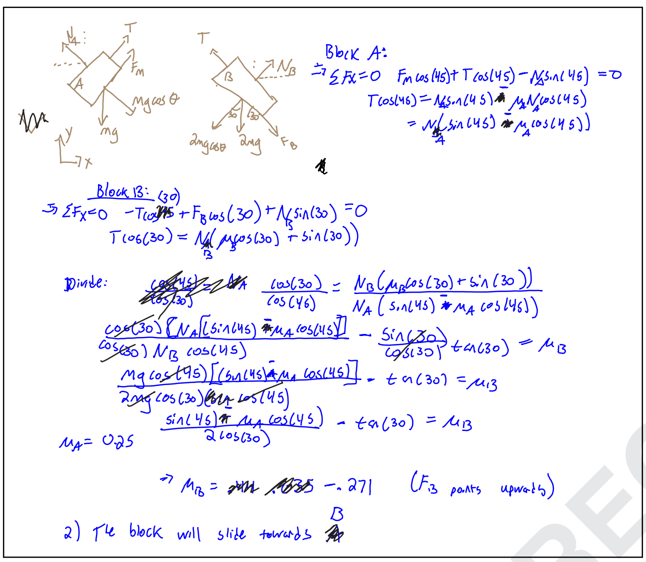

Fig. 2. A solution to the Statics problem from Fig. 1. Brown = free body diagram, blue = equation, and black = cross-out. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

types of features and train classifiers to predict the grade achieved on the work. For example, temporal features describe the order in which problem-solving activities are performed and the amount of time spent on each. Spatial features describe the extent to which a student revisits earlier parts of a solution to revise the work. To construct their features, Van Arsdale and Stahovich require that the pen strokes be separated into free body diagram, equation, and cross-out strokes. Our techniques provide an automated means of doing this.

Our stroke classification method builds upon methods for single stroke classification developed by Blagojevic et al. [6] and Stahovich et al. [46]. Both of these methods are intended to be general-purpose and were evaluated on small datasets collected in the laboratory. By contrast, our method is designed for handwritten problem solutions and has been used on large datasets of students' coursework.

Similarly, our grouping methods build upon the work of Stahovich et al. [46]. Just as they do, we use a pairwise classifier to determine if pairs of pen strokes belong to the same objects. However, while their classifier uses features intended to be general-purpose, we use features that are optimized for handwritten equations.

Our work makes several important contributions. First, we developed an effective method for the three-way classification of cross-out, free body diagram, and equation strokes. This is a challenging problem because cross-outs comprise a very small minority of the pen strokes. Our stroke classification approach is more accurate than prior stroke classification methods for our task. Second, we developed a method for identifying equation groups in handwritten problem solutions. Our equation grouper is more accurate than prior stroke grouping methods for this task. Third, we developed a character grouper that is as accurate as prior methods, but is better suited to our task as it uses features computed from quantities already computed for our equation grouper. Fourth, our set of methods makes it possible to perform data mining of handwritten coursework on a large scale. Fifth, we have demonstrated the usefulness of our methods for a novel application in educational data mining.

2. Related work

2.1. Single stroke classification

There are several existing techniques for single-stroke classification. Most are designed for distinguishing handwritten text from shapes (i.e., non-text). For example, Jain et al. [22] use stroke length and stroke curvature to distinguish text from non-text in online documents. Qi et al. [35] present a method for using conditional random fields to classify strokes in organizational chart diagrams as either connectors or boxes. Addressing a similar problem, Bishop et al. [5], Patel et al. [33], and Bhat and Hammond [4] present methods that integrate shape and temporal information for classifying individual strokes as either text or drawing strokes. Wang et al. [53] improve on Bishop et al.'s method. Indermühle et al. [21] distinguish text from non-text using a top-down approach that segments a document into regions of text and non-text. They also developed a bottom-up approach that examines the neighborhoods of individual pixels and considers connected components. More recently, Delaye and Liu [9] used conditional-random fields to jointly model local, spatial, and temporal information to achieve accurate discrimination of text from non-text.

Our stroke labeling approach is similar to approaches developed by Blagojevic et al. [6] and Stahovich et al. [46]. Both of those methods characterize pen strokes with a set of features and then use a classifier to determine the semantic class of a stroke. These methods employ features that are intended to be general-purpose. By contrast, we have developed a set of features optimized for handwritten coursework. Also, we have developed a novel two-classifier approach to achieve high accuracy on cross-out strokes, which comprise only a small minority of the pen strokes. The first classifier distinguishes cross-outs from non-cross-outs. The second differentiates the latter into free body diagram and equation strokes. As shown in Section 6, our method is more accurate than the methods in [6,46] for our task.

2.2. Stroke grouping

Grouping pen strokes into objects is a challenging problem. Many existing recognition systems avoid this problem by placing constraints on the way users draw. For example, some systems require the user to provide explicit cues, such as button clicks or pauses, to demarcate each object [19]; others require each symbol to be drawn with a single-stroke [41,55,38,17] or a temporally contiguous sequence of strokes [12]. While these constraints aid recognition, they do not generally match the way people naturally draw [2].

Some automated grouping methods rely on search. These methods consider a large search space of possible stroke groups and use recognition to identify which groups represent objects. For example, Shilman and Viola [45] used A* search to generate candidate groupings and then evaluated them with a recognizer to

Please cite this article as: Stahovich TF, Lin H. Enabling data mining of handwritten coursework. Comput Graph (2016), http://dx.doi.org/10.1016/j.cag.2016.01.002