12-1-2008

Ballot Mark Detection

Elisa H. Barney Smith
Boise State University

Daniel Lopresti
Lehigh University

George Nagy
Rensselaer Polytechnic Institute

Ballot Mark Detection

Elisa H. Barney Smith[1], Daniel Lopresti[2], George Nagy[3]
[1]*Boise State University,* [2]*Lehigh University,* [3]*Rensselaer Polytechnic Institute,*
EBarneySmith@boisestate.edu, lopresti@cse.lehigh.edu, nagy@ecse.rpi.edu

Abstract

Optical mark sensing, i.e., detecting whether a "bubble" has been filled in, may seem straightforward. However, on US election ballots the shape, intensity, size and position of the marks, while specified, are highly variable due to a diverse electorate. The ballots may be produced and scanned by poorly maintained equipment. Yet near-perfect results are required. To improve the current technology, which has been subject to criticism, components of a process for identifying marks on an optical sense ballot are evaluated. When marked synthetic ballots are compared to an unmarked ballot, the absolute difference of adaptive thresholded images gives best detection rates for all darknesses of marks, but at a false alarm rate increase. Simple absolute differencing can give good detection results with lower false alarm rates.

1. Introduction

In the wake of disputed US elections in 2000, Congress provided funds for new election machinery through the Help America Vote Act (HAVA). Because of growing dissatisfaction with touch-screen displays, many election districts are now leaning towards optical mark sensing equipment for processing paper ballots.

A ballot consists of a set of *contests*. A contest may select a single candidate for an office from among several candidates (for instance, for State Senator, or City Comptroller), several candidates (e.g., 6 City Council Members from 15 candidates), or offer a binary choice (retention/dismissal of a judge or adoption/rejection of a statutory proposition).

A ballot comprises *identification* (election district, date of election, ballot style number, page number), *instructions* (to select a candidate, correct mistakes, and cast the vote), a list of *contests* and *alternatives* (candidates' names and party affiliation for political offices, propositions), and a set of *targets* to be marked for each vote. Each voter places a *mark* in the appropriate targets. Then the voter either runs the ballot through a scanner (sometimes called a Portable Ballot Counter or PBC), or puts it in an envelope for processing at election headquarters after voting closes.

The instructions on the ballot usually specify what constitutes a valid mark (e.g., "darken the oval completely with a #2 pencil or black pen"). In contrast to the "bubble" answer sheets used for standardized exams, what determines the legal validity of the interpretation of a particular ballot is the *voter's intent*. In many states, election officials affiliated with the competing parties work in teams to asses mark validity.

We evaluate the capability of different algorithms to distinguish marks from registration noise and explore the accuracy and consistency of automated image processing under various scenarios. The metrics we investigate are the percentage of *detected spurious marks* (false positives) and of *missed marks* (false negatives) as a function of the *size* and *contrast* of the marks and of the effect of marks *overlaying* text or graphics.

We experiment with synthetic ballots because we will eventually need an extreme range of mark variation to address problems at the tail of the curve. We report results on synthetic optical sense ballots produced by placing marks with controlled variations on images of real blank ballots (which also avoids tedious manual mark characterization). We hope that our results from these and future experiments will help improve (1) ballot design, (2) optical mark sense hardware and software, and, ultimately, (3) definition of what constitutes an *intentional mark*. Although optical mark sense technology debuted 80 years ago, we are not aware of any comparable published research.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

2. Ballot and Mark Data

We experimented with a base ballot template from Minnesota (Fig. 1) that is generally representative of the filled-oval format of ballots. Although the instructions for this ballot specify that the voter should fill in the oval targets, some voters may well use a check mark or an X. Voters may also drag their pencil and leave stray strokes or *hesitation marks* (small dark dots). Our approach is designed to detect of all of these mark types. Later analysis can be designed to distinguish between mark types.

Synthetically marked ballots were created using the methods described in [3]. Four marked ballots were evaluated for these tests, with 58 marks each, most intended to be "difficult." Marks were entered with a variety of shapes (oval, dot, check and X), five gray intensity levels, five sizes and a variety of positions relative to the target ovals (Fig. 2). Most marks are centered in the target oval, but 37% are displaced far enough to overlap the candidate or party text, or the ballot rulings. While in a real election there should be only one vote for each office, we deliberately applied marks to all the targets in order to reduce image file handling. We have placed these ballots online [4].

3. Mark Processing Methods

Mark detection requires more than determining whether the content of a target position exceeds a threshold. The presence of possibly valid marks outside the nominal positions brings ballot image processing from optical mark sensing to a variation on forms reading.

In traditional forms-processing, the material of interest will be (predominantly) found in specific fields. Accuracy is improved by the use of context from a database: pre-recorded names, addresses and part numbers. Processing then involves recognizing and registering the form, and extracting the new text using context. In contrast, the voter is anonymous, and we cannot even use priors like "most voters in Idaho vote Republican." As in forms, a ballot enrollment stage identifies the locations of the target ovals. Ideally the marks will consist of large solid black marks and only the presence or lack of sufficient fill within the oval would need to be detected. In reality even marks that do not follow the ballot instructions must be located and identified because the legal definition of a vote is *voter intent*. In some jurisdictions, any of the marks shown in Figure 2 that appear somewhere within a target would be considered valid votes, while marks that appear completely outside the target area are important to detect so that the ballot can be flagged for followup examination.

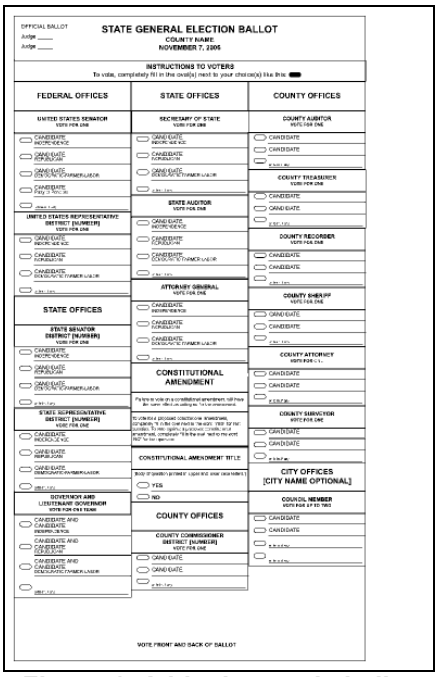

Figure 1. A blank sample ballot

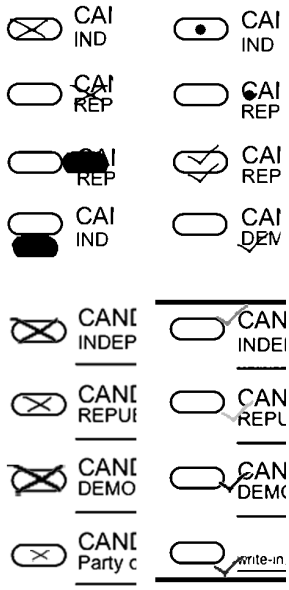

Figure 2. Part of a synthetically filled ballot with examples of four mark types, four alternate sizes and four alternate gray levels.

If ballots were printed with drop-out ink, then only user-added marks would be recorded when the page was imaged with the appropriate light. Since most election districts are not willing to accept this additional cost, the form background must be removed through image registration and image differencing. The blank template ballot image is aligned with the marked ballot image using a frequency based correlation filter [5]. The image

difference is calculated to reveal the added material in a manner similar to forms processing [8, 10].

Five different image differencing algorithms are evaluated while looking for additions to the blank ballot. Even with perfect image registration, the noise from variations in printing and scanning will introduce differences. We report how each of these differencing techniques performs in the context of identifying marks of unknown shapes in arbitrary positions.

The five differencing techniques are as follows:

D1. The absolute value of the difference between the raw images.
D2. The absolute value of the difference between the raw images, smoothed by a 3x3 uniform kernel.
D3. The absolute value of the difference between 3x3 smoothed images.
D4. The absolute value of the difference between 3x3 smoothed images, smoothed by a 3x3 uniform kernel.
D5. The absolute difference of smoothed adaptively binarized images.

A median filter was applied in all cases to reduce the effects of spatial sampling phase and additive noise. After thresholding the difference image, connected components were identified. Components smaller than 2x2 pixels were removed.

Morphological closing with a disk of radius 10 was then applied This was intended to merge components that were broken because the marks crossed text or rulings. It also mended some of the X and check marks that had split because the smoothing had lowered the intensity of part of the mark stroke below the threshold.

Experimental Results

The filtered difference images were thresholded at threshold levels ranging from 96 to 240 (on a 0 to 255 scale) in steps of 8. A lower threshold retains more of the difference and therefore fewer marks are missed. A higher threshold increases the false alarm rate. With thresholds up to 152, all black mark additions were retained with every differencing method. Differencing method D2 was the most sensitive to the choice of threshold level. The checks and X's were more sensitive to the threshold than the filled ovals and dots in methods D2 and D4, and less in methods D3 and D5. On bilevel ballot images, method D1 was not sensitive to threshold.

Table 1 shows the base results with a threshold of 120 over all ballot samples including a range of mark shapes, contrasts and sizes. Very few false alarms (#FA) were detected in methods D1-D4. In method D5 the morphological closing increased the size of many of the adaptive threshold ghosts beyond the 2x2 cutoff, so more of them were detected. Several of the false alarms related to the same ballot image defects were spatially clustered. Some of the missed check marks occurred because some neighboring check marks, as shown in Fig. 2, were merged during the closing and counted as only a single detection.

Since not all marks will be made with black pen or medium soft pencil, the data set included marks with a range of different gray levels: {0, 36, 80, 132, 190}, where 0 is black, and 255 is white. Differencing method D5 had the best detection accuracy for all mark intensities due to its use of adaptive thresholding. Adaptive binarization allows the differencing threshold to be decreased without missing more marks. A differencing threshold of 144 yielded 100% detection with no false alarms. The effect of gray level on the detectability depended on the mark shape. At low thresholds, 100% of the ovals can be detected with every differencing method. The errors in the gray=0 case are due to the merged marks reported earlier.

The other variable in the marks in our dataset was the size of the mark. The marks had a base size, and a subset of the marks in the data set were produced in sizes 50%, 75%, 125% and 150% of that base size. The size of the mark had no effect on the detectability on this data set.

Conclusion

In our experiments the color or darkness of the mark was the biggest factor in the detectability of the marks. Marks were best detected when adaptive thresholding was used, but this lead to a significantly higher false alarm rate, and requires more processing time. The size and shape of the mark did not affect the detection performance. The choice of threshold is

Table 1: Detection rates by mark shape over entire data set of 232 marks .

	# FA	% Detected Total	% Detected X	% detected Check	% Detected Oval	% Detected Dot
D1	**3**	**97.4**	**97.0**	**95.8**	**100.0**	**98.1**
D2	**4**	**95.7**	**94.0**	**94.4**	**100.0**	**96.2**
D3	**7**	**96.6**	**97.0**	**94.4**	**100.0**	**96.2**
D4	**5**	**94.4**	**97.0**	**88.7**	**97.6**	**96.2**
D5	**31**	**99.1**	**100.0**	**97.2**	**100.0**	**100.0**

Table 2: Detection rates given mark gray level intensity.

	Mark Gray =0	Mark Gray =36	Mark Gray =80	Mark Gray =132	Mark Gray =191
D1	**99.5**	**100.0**	**100.0**	**100.0**	**63.6**
D2	**100.0**	**100.0**	**100.0**	**66.7**	**45.4**
D3	**99.5**	**100.0**	**100.0**	**100.0**	**45.4**
D4	**99.5**	**100.0**	**100.0**	**58.3**	**45.4**
D5	**99.5**	**100.0**	**100.0**	**100.0**	**100.0**

important, as is the size of the structuring element for morphological closing. A smaller structuring element would not merge adjacent marks, but more detected marks would be broken. Only a few false alarms occurred. Identifying marks for which additional logic is necessary was one of the goals of these pilot experiments.

Only translation between the template ballot and the test ballot was accommodated in our experiments. Skew or scale distortion, such as often found in scanned ballots (with skew being more likely than scale), can be estimated and corrected with the Fourier-Mellon transformation [2, 6].

With a wider range of ballot images, the false alarm rate will increase. Here only differences were identified. To distinguish marks from noise the identification of the detected components could be supplemented by exploiting the expected consistency of marks on each ballot. We could compare each detected mark candidate to the average or median of all the marks on the same ballot. If most of the marks are large checkmarks, then it would be reasonable to classify a small X as a hesitation mark. On the other hand, if the majority of the marks consist of a small X, then a large oval might be an extraneous blob or erasure. The mathematical framework for this kind of analysis, dubbed *style*, appears in [7, 9].

The above characterization, applied to real ballots, may be sufficient not only to establish an algorithm for detecting the marks, but also to determine the validity of a ballot and the resulting tally. What types of marks are acceptable, and how much variation among individual marks on a single ballot can be tolerated, must of course be left to election officials. We can, however, simulate various scenarios, and compare the results on synthetic ballots (prepared to mimic the distribution of marks expect in actual elections) with those obtained by submitting the same ballots to commercial optical-sense ballot counting devices.

The development of an ultra-reliable and trustworthy paper-based voting technology would have broad impact. Such technologies tend to win acceptance slowly. Right now, however, we are at a cross-roads, with several radically different voting technologies competing for acceptance. It is therefore timely to direct attention toward the role that paper records can play. We hope our work will inspire the research community to take a closer look at some of the interesting technical problems that arise.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grants No. CNS-0716647 (EBS), CNS-0716393 (GN) and CNS-0716368 (DL). The synthetic ballots were constructed by Anne Miller.

References

[1] R. C. Gonzales, R. E. Woods, *Digital Image Processing*, Addison-Wesley, Boston, Massachusetts, 1992.

[2] L. A. D. Hutchison, W. A. Barrett, "Fast Registration of Tabular Document Images Using the Fourier-Mellin Transform," *Proc. Document Image Analysis for Libraries*, Palo Alto, California, January 2004, pp. 253-267.

[3] D. Lopresti, G. Nagy, and E. H. Barney Smith, "A Document Analysis System for Supporting Electronic Voting Research," *Proc. Document Analysis Systems*, Nara, Japan, September 2008.

[4] The PERFECT Project: RPI Synthetic Ballots http://perfect.cse.lehigh.edu/BallotTestData_RPISyntheticBallots.html

[5] W. K. Pratt, *Digital Image Processing*, John Wiley & Sons, New York, 1991.

[6] B. S. Reddy, B. N. Chatterji, "An FFT-Based Technique for Translation, Rotation, and Scale-Invariant Image Registration," *Trans. Image Processing*, Vol. 5, No. 8, August 1996, pp. 1266-1271.

[7] P. Sarkar, G. Nagy, Style consistent classification of isogenous patterns, *IEEE Trans. PAMI-27*, 1, pp. 88-98, January 2005.

[8] S. L. Taylor, R. Fritzson, "Registration and region extraction of data from forms," *Proc. ICPR*, 1992, pp. 173-176.

[9] S. Veeramachaneni, G. Nagy, "Analytical results on style-constrained Bayesian classification of pattern fields," *IEEE Trans. PAMI-29*, 7, pp. 1280-1285, July 2007.

[10] B. Yu, A. K. Jain, "A generic system for form dropout," *IEEE Trans. PAMI*-18, 11, 1127-1134, 1996.