NISTIR 8069

Systematic Measurement of Marginal Mark Types on Voting Ballots

Andrea Bajcsy
Ya-Shian Li-Baboud
Mary Brady

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8069

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8069

Systematic Measurement of Marginal Mark Types on Voting Ballots

Andrea Bajcsy
Ya-Shian Li-Baboud
Mary Brady
Software and Systems Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8069

July 2015

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

Abstract

The presence of marginal marks on voting ballots is a known problem in voting systems and has been a source of dispute during federal and state-level elections. As of today, marginal marks are neither clearly countable as votes or as non-votes by optical mark scanners. We aim to establish quantitative measurements of marginal marks in order to provide an objective classification of ballot-mark types and ultimately improve algorithms in optical scanners. By utilizing 800 publicly available manually-marked ballot image scans from the 2009 Humboldt County, California election, we established a set of unique image features that distinguish between votes, non-votes, and five marginal mark types (check-mark, cross, partially filled, overfilled, lightly filled). The image features are related to semantic labels through both unsupervised and supervised machine-learning methods. We demonstrate the feasibility of developing an automated and quantifiable set of custom features to improve marginal mark accuracy by 4 to 8 percent, depending on classification model.

1. INTRODUCTION

The use of optical scan paper ballot systems has been on the rise and they are among the most widely used electronic voting equipment in the United States [1]. In 2012, 56 % of voters used a paper ballot with an optical scan machine during the presidential election [2]. Optical scan paper ballots provide a voter-verifiable paper audit trail. Optical mark scanners have improved in cost and reliability during recent years and brought about the resurgence of paper ballot voting within the United States [3]. Mark-sense ballots contain predefined voting targets, typically in an oval or broken arrow format, intended for a voter's mark. To process mark-sense ballots, the voting system uses image processing techniques to detect votes. For example, after a vote is placed on a mark-sense ballot, commercial optical mark scanners count the number of dark pixels in a voting target to determine whether a vote was or was not placed by the user [4]. Regardless of whether the mark-detection algorithms rely on darkened pixel counts or edge detection, there is still an open problem of detecting non-typical voting marks through image analysis.

Voter-generated ballot markings often vary from the ideal mark that users are instructed to produce. In the universe of all possible marks, marks can be categorized into either *reliably detected*, or *marginal marks* [4]. In the context of mark-sense ballots, *reliably ignored marks* are extraneous, never count as votes, and include empty voting targets and hesitation marks. *Reliably detectable* marks are always counted as votes and include mostly filled voting targets and check marks, although scanners do exhibit variation in their ability to detect such marks [4]. Finally, *marginal marks* may or may not be detected by the mark scanner and include small lines, atypical ink or marker, and marks outside but near the voting target (Figure 1).

Figure 1. Examples of marginal marks from our dataset. We establish five baseline classes of marginal mark types based on the most commonly occuring atypical marks, such as crosses (a,b), lightly filled (c), and partially filled (d) marks seen above.

The Election Assistance Commission (EAC) Draft Voluntary Voting System Guidelines (VVSG) 1.1 defines *marginal marks* as "neither clearly countable as a vote nor clearly countable as a non-vote" [5].

Since marginal marks are not clearly distinguishable by optical mark scanners, there is potential for such marks to be inconsistently counted through multiple runs within a scanner or across multiple scanners. The inconsistency can be attributed to the physical properties of the sensor where no two sensors are identical and the behavior can further be affected by the environmental conditions. Thus, there is a **need to develop systematic quantitative measurements and models of marginal marks** that can be used by all optical mark scanners for classification, testing and calibration in an effort to enable consistent and transparent counts of votes. Furthermore, the quantitative features and models can also be used to provide a systematic, unbiased and transparent means of auditing optical scan ballot machines by doing an independent count and extracting problematic ballots that may cause discrepancies in the vote count. Automated extraction of quantitative features and well-characterized models also contributes to the ability to address the EAC Draft VVSG 1.1 requirements, including the need for ensuring repeatability and a voting system that does not introduce bias [5] [6].

One of the basic tenets of the U.S. election system is ensuring that every vote is counted. The existence of marginal marks presents a challenge to this tenet, as marginal marks can often be miscounted or ignored. Determining the type of marginal mark and whether it is a vote or non-vote requires both the judgment of human assessors and the reproducible, unbiased classification by optical mark scanners. Contested elections, such as the 2008 U.S. Senate race in Minnesota and 2008 New Hampshire Democratic Primary, highlight the disagreement between the machine and human interpretations of a ballot mark and the potential for a miscount of votes [7]. In today's digital era, machine interpretation is understood as an automated classification algorithm operating on quantitative measurements of ballot marks. Therefore, there is a **need to relate automated classification and human interpretation of ballot marks**. Previous work regarding this need has focused on creating efficient user-guided ballot image verification systems [8] and on analysis of write-in votes on ballots [9]. These approaches have expanded the application of image processing techniques to the ballot mark variability problem. However, little research exists on determining the subcategories of marginal mark types and the potential inconsistency in interpretation by both machines/algorithms and humans.

1.1 Problem Statement

Based on the existing literature and previous research, systematic categorization of marks that fall outside of the ideal vote and non-vote categories (e.g., marginal marks) is an open issue. To address this issue, the following overarching research question was posed: *How can marginal mark types be automatically classified into human-understandable categories?*

The need to answer the research question is based on various state codes when it comes to whether a mark is recognized as a vote or no-vote. Some state codes require both prescribed and acceptable marks to be counted [10]. Examples of acceptable marks include, but are not limited to, checks and crosses if the marks are within a defined target [11].

The overarching research question can futher divided into three sub-problems:

(1) What mark-image features are necessary for classifying marginal mark types?

(2) How do we establish ground-truth for human-understandable mark categories?

(3) What classification model is the most accurate for marginal mark-type classification?

Note that we do not make any assumptions about voter intent nor do we attempt to determine if vote marks should count as a vote or non-vote. The approach is to provide quantitative characterization of mark types to enable manufacturers and states to have an unbiased and repeatable means of specifying acceptable mark types.

Given a set of ballot images, the goal was to demonstrate the feasibility of defining quantitative measurements of ballot mark images and relate them to human assessments of representative mark types (e.g. check, cross, partially filled, lightly filled, overfilled). This was accomplished by linking the algorithmic model interpretation of vote marks with human interpretation (which we assume to be the ground truth) in order to assure the consistency of automated marginal mark classification. The resulting set of quantitative features and models serves as a means to mathematically define marginal marks and common mark types. The characterization enables a means to benchmark optical scanner image processing algorithms to ensure transparency and consistency in mark interpretation.

2. BALLOT IMAGE DATASET

Our input data set was taken from the Humboldt County, California May 2009 election [12]. The total dataset contains 7.4 GB of 26000 publicly available voting ballot image scans in JPEG format. Each 1265 pixel by 1648 pixel scanned ballot image has a resolution of 72 dpi. However, for the purposes of demonstrating feasibility, we utilized the first 800 ballot images in the dataset for initial training and testing our classification algorithms. Another 800, for a total of 1600 ballot images, were processed as a future test set. A subset of available images were chosen to focus on the design of an automated mark classification pipeline and minimize the computationally intensive image processing time

Our algorithms take advantage of several unique features present on the ballots from our dataset (Figure 2). Each ballot has an ideal vote mark printed at the top of the ballot and has borderline tick-marks originally used for the optical mark scanner calibration printed along the left side. We note that if other ballots are being used, then this portion of the process will need to be modified according to the new ballot features. We also created a set of human-understandable mark categories which represent and distinguish mark type patterns within the data set. In our case, these mark categories consist of empty, filled, lightly filled, partially filled, overfilled, cross, and check marks. Finally, we make no assumptions about the voter's intent when marking the ballot.

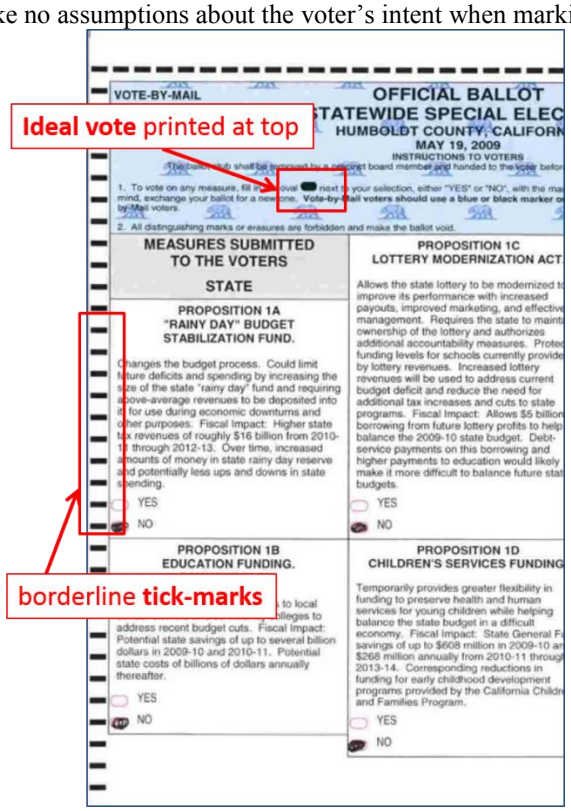

Figure 2. Portion of a Humboldt, California 2009 Election Ballot with borderline tick-marks and ideal vote assumptions.

3. AUTOMATED CLASSIFICATION OF BALLOT-MARK TYPES

In order to address the three main questions listed in Section 1.1, we developed a processing pipeline for automating classification of ballot-mark types (see Figure 3):

(1) *Image Preprocessing and Ideal Vote and Ideal Non-Vote Extraction*
Registration of ballot image scans to common coordinates, generation of ideal vote and non-vote ground truth, and extraction of mark region of interest.

(2) *Unsupervised Learning*
Computation of each extracted mark's correlation to ideal vote and ideal non-vote, establishment of threshold for marginal mark detection.

(3) *Crowdsourcing: Human Marginal Marks Assessment*
Design, implement and use of web-based application for semantic label collection from human assessors, filtering of marks that are considered filled by the majority of assessors, where "majority" is defined in section 3.3.

(4) *Supervised Learning*
Supervised learning is a machine learning technique using a set of measured features and training data, which has an associated ground truth, to train a classification model based on a learning algorithm. The mark type defined by the majority of the human assessors serves as the ground truth in this study. The training set is a set of marginal marks with established ground truth. Each mark underwent automated extraction of 33 off-the-shelf image

features, 22 baseline image features, and 9 custom features, and comparison of decision and non-decision tree classification models for each feature set.

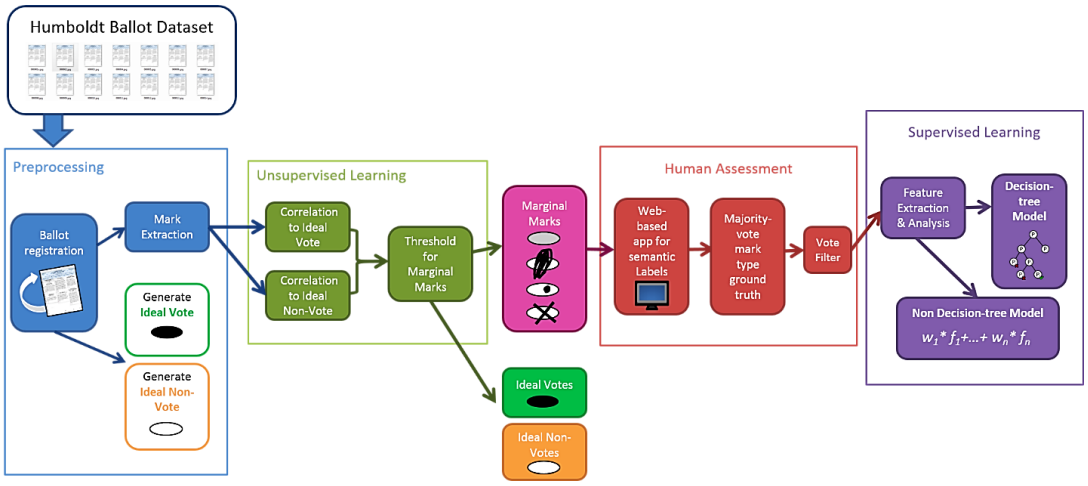

Figure 3. Processing pipeline for automating classification of ballot-mark types

3.1 Image Preprocessing and Ideal Vote and Ideal Non-Vote Extraction

The preprocessing of ballot images involved registering the ballots to a common coordinate system and generating ground truth images for vote and non-vote marks.

3.1.1 Image Registration with Template Ballot

Each ballot image scan in our dataset is not guaranteed to be perfectly aligned and may exhibit a varying degree of rotation or translation. These variations affect both the accuracy of our ballot mark extraction and the consistency of our image metrics. Thus, we register each ballot scan image to an ideal ballot template present within the dataset. The template ballot is chosen through visual inspection and, in the case of the 800 image training ballot set, was the first ballot present in the set. To perform the registration, we utilized the open-source Fiji[1] application's rigid registration with a similarity feature extraction model. One problematic image (e.g., partially scanned) in the data set degraded the registration of all the subsequent ballot images. To remedy this, we performed a quick visual scan of the input dataset and removed any empty or partially scanned images (about 3 images from 1600 images comprising training and testing data). It is important to note the need for a robust registration process that can detect or gracefully handle problematic images in a manner that preserves accurate registration of all valid images in the data set.

3.1.2 Ideal Vote and Ideal Non-Vote Extraction

In order to measure the deviation of an extracted mark from an ideal mark, we needed to generate ground truth markings for comparison. Theoretically, the only established ground truth for a voting mark is an ideal vote and an ideal non-vote. We considered either creating simulations of an ideal vote marking and an ideal non-vote marking, or computing the measured average of the printed

[1] *Certain software are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.*

ideal vote and ideal non-vote. We opted to compute the printed average in order to capture the printing and scanning variation from one ballot image to the next.

To generate the ideal vote ground truth, we relied on the assumption that each ballot image had the ideal vote printed at the top of the ballot. For each ballot scan within our dataset, we extracted the ideal vote based on its consistent location at the top of the image. In order to create the composite of all ideal vote marks, we computed the average of each pixel across all the ideal vote mark sub regions. Then, to eliminate extraneous text and noise, we performed a color threshold on the composite ideal vote image and resulted in our final baseline ideal vote (see Figure 4).

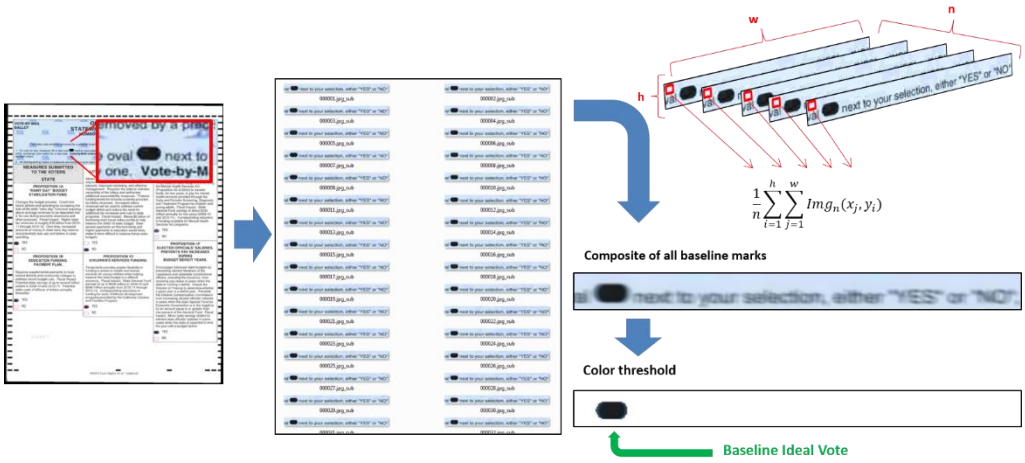

Figure 4. Ideal Vote Extraction and Generation

A similar process was employed for the generation of the baseline ideal non-vote. However, in order to extract each of the non-vote marks within a ballot, first all the mark regions of interest were extracted. Since non-vote marks have a higher occurrence of white pixels than black pixels compared to vote marks, extracted marks that fell within a threshold of color distribution were considered to be empty and visually inspected for any anomalies.

3.1.3 Ballot Mark Region Detection and Mark Extraction

We defined a mark region of interest as a 380 pixel by 30 pixel region which captures as much of the user's mark in and around the mark target as possible based on our visual inspections of ballots with external marks from the ellipse. This was the maximum size that the region of interest (ROI) could be without encroaching on a neighboring mark's region. The ROI is considered the mark region. In order to extract each of the marks, we analyzed the border tick-marks around the left edge of the ballot image scans as guides to the locations of our ROI's. A lookup table of the desired tick-mark locations enabled us to crop consistent mark regions and save each extracted mark into a separate dataset for further study.

3.2 Unsupervised Learning to Detect Marginal Marks

In order to separate marginal marks from those that are reliably detectable (vote) or reliably ignored (non-vote), we used a combination of normalized cross-correlation and clustering analysis. Each mark that was extracted from the ballot image dataset was correlated to the ideal vote and ideal non-vote images generated during the preprocessing step. We chose normalized cross-correlation as a similarity metric due to its robustness to varying lighting and exposure conditions [13]. The formula for computing normalized cross correlation is shown in Eq. (1).

$$NormCorr(x,y) = \frac{1}{numPixels_f + numPixels_t} * \sum \frac{(f(x,y)-\mu_f)(t(x,y)-\mu_t)}{\sigma_f * \sigma_t} \quad (1)$$

Where f is a mark image and t is an ideal vote or an ideal non-vote image, μ_f and σ_f are the average pixel intensity and standard deviation respectively of f, μ_t and σ_t are the average pixel intensity and

standard deviation, respectively of *t*. The abbreviation *numPixels* refers to the total number of pixels within an image and *x* and *y* represent the row and column locations of a pixel within an image.

The results of computing the correlation for each mark image with respect to the ideal vote and ideal non-vote images are summarized in Figure 5. As expected, two predominant clusters formed, dividing the empty marks from those that were filled. The similarity metric used appears to be sensitive to variations in the appearance of target labels, registration marks, as well as rotation and translation of the ROI as demonstrated by the generally low correlation coefficient to non-votes. The cluster generated from the filled marks also exhibited greater overall spread than the empty marks. Upon closer inspection, the correlation values of marks that fell in between the two clusters represented many of the marginal mark cases (such as lightly filled marks, checks, or strikethroughs). Thus, empirical threshold values were used to delineate the marginal marks from the vote and non-vote marks in Figure 5 according to Eq. (2).

$$0.21 \leq normCorr(f, t_{vote}) \leq 0.62 \;\; ; 0 \leq normCorr(f, t_{non-vote}) \leq 0.45 \qquad (2)$$

These threshold values were designed to be lenient in order to keep as many atypical marks as possible. This analysis resulted in 461 marginal marks, 4984 non-votes, and 4431 votes. The marginal marks constituted about 4.9 % of total ballot marks. Only the 461 marginal marks continued to the Web-based human assessment portion of our pipeline. While improvements can be made in the similarity metric to be invariant to rotation, translation and standard ballot markings, the results from the normalized cross-correlation resulted in the ability to use quantitative thresholds to extract marginal marks from the reliably detectable votes and non-votes.

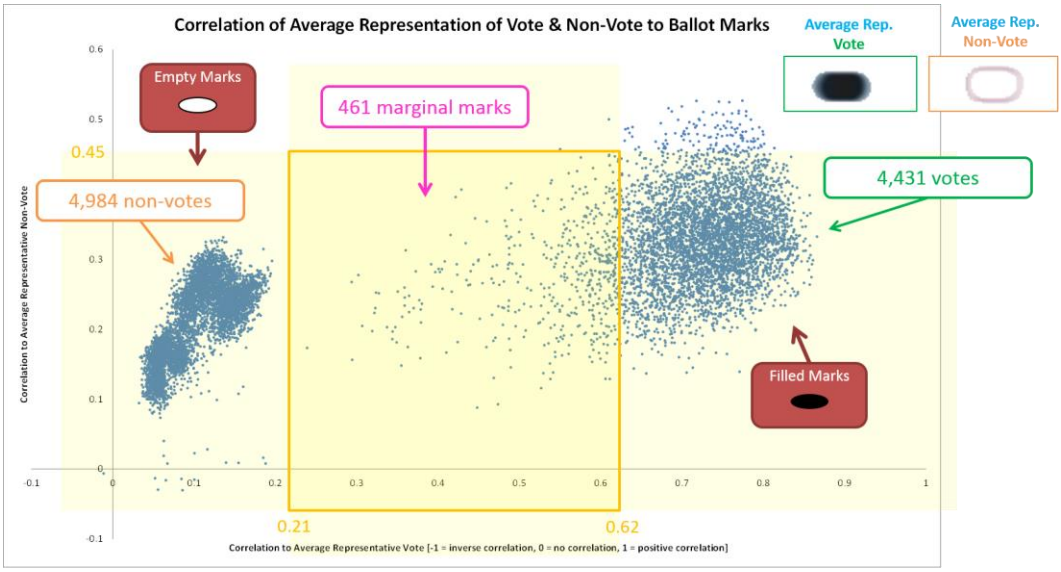

Figure 5. Unsupervised learning threshold for marginal marks, vote marks, and non-vote marks.

3.3 Web-based Human Assessment of Marginal Mark Types

In order to provide a crowdsourcing infrastructure to gather semantic labels, or descriptions of mark categories, for each of the 461 marginal marks, we designed a web-based assessment to collect human interpretation of marginal mark types. The website was constructed using Hyper-Text Markup Language (HTML) and Cascading Style Sheets (CSS) and JavaScript. All of the user-provided information about marginal mark types was stored in a MySQL database. The EAC's definition of a marginal mark and was provided with the following instructions:

- You will be given a mark extracted from a ballot.
- Classify the mark as *Marginal* or *Not Marginal*
- Classify the type of the mark based on the given types, or click *Other* and submit your own type.

- You may stop at any time and return to the same mark where you left off.

After reading the instructions, the assessor was presented with a single mark image and seven mark types to choose from (see Figure 6). Each semantic label (check, cross, partially, lightly, overfilled, filled, empty) was accompanied with a synthetically created visual representation of the mark type to reduce interpretation variation between assessors. By gathering human classification of the 461 marks, we were able to generate a ground-truth mark type for each extracted mark based on a majority vote. However, consensus on each extracted mark's type was not always unanimous, so we performed vote-based filtering on three criteria:

1. **"Tie" marks** are marks with two or more types receiving the same amount of votes.
2. **"Fusion" marks** are marks with < 87.5 % agreement on a mark type. For establishing ground truth, ideally we would have 100 % consensus. However, given the limited number of samples for certain mark types the 87.5 % threshold was chosen to ensure sufficient data was available for each mark type. Eight of the ten assessors completed all 461 marks for the ground truth assessment. Therefore the threshold means only one assessor did not agree.
3. **"Filled" marks** are where the majority voted for filled.

"Tie", "Fusion", and "Filled" marks were removed from the training dataset because either the human assessors could not reach a consensus about the mark type, or the marks were considered filled and therefore did not comply with the definition of marginal mark (see Table 1). After post-filtering 461 mark images we obtained 168 images with ground truth labels.

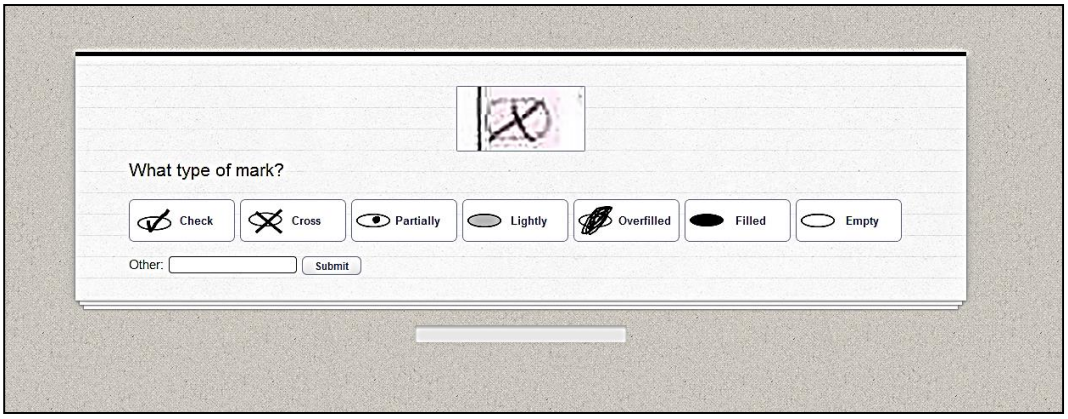

Figure 6. Web interface for human-assessment mark type data collection

One final analysis was performed on the human-generated mark types to relate semantic categories of mark types and classes of marginality according to human assessors. Table 1 summarizes the relationship. According to **Table 1**, there were 168 marks labeled as marginal and falling into one of the five semantic categories (cross, check, partially, lightly, and overfilled).

Table 1. Relationship between semantic categories of mark types (columns) and classes of marginality (rows)

Mark Type / Marginal?	Cross (50)	Check (7)	Partially (35)	Lightly (50)	Overfilled (26)	Filled (39)	Fusion (232)	Tie (22)
Marginal	47	7	34	47	23	1	115	17
Non-Marginal	2	0	1	1	0	38	84	1
Tie	1	0	0	2	3	0	33	4

Five mark types are mostly seen as marginal

Because "Fusion" marks may consist of 2+ other mark types, they are harder to classify as marginal or non-marginal

3.4 Supervised Learning to Relate Image Features with Marginal Mark Types

Using the mark-type labels collected from the 10 human assessors, we were able to create a labeled training dataset. In addition, we needed to relate the mark-types chosen by humans with mark image features in order to develop a mark-type prediction function. Three image feature sets were evaluated for the automated classification of marginal mark types. The first set was the off-the-shelf Fiji application's 33 image features (e.g., area of the ROI, standard deviation of pixel intensity). The images were binarized using the Fiji application's automatic threshold function. The 33 built-in features were each extracted from binarized mark images. We designed our second image feature set to capture 22 baseline patterns such as average intensity inside mark target ellipse, correlation to ideal vote, and number of black pixels in each quadrant of the ROI. The objective of our design was to better characterize each mark type by incorporating a priori known salient attributes of the ballot and each mark type.

Finally, the third group of 9 custom image features was specifically designed to focus on distinguishing one mark type from the rest as described in Table 2.

Table 2. Description of the custom features for specific mark types.

Feature	*Formula*
Partially Filled	$v_{bin} = \frac{Black\ pixels\ inside\ mark\ target \cap outside\ mark\ bounding\ box}{Black\ pixels\ inside\ mark\ target \cap inside\ mark\ bounding\ box}$ Partially filled marks tend to have a minimal amount of dark pixels outside of the ellipse. The also tend to have a contrast of white and darker pixels inside the mark target region which can be captured by the horizontal, vertical and diagonal boxes.
Checks & Crosses	$v_{binLR} = Black\ pixels\ on\ left\ to\ right\ diagonal\ of\ mark\ bounding\ box$ $v_{binRL} = Black\ pixels\ on\ right\ to\ left\ diagonal\ of\ mark\ bounding\ box$ $v_{binV} = Black\ pixels\ on\ vertical\ center\ of\ mark\ bounding\ box$ $v_{binH} = Black\ pixels\ on\ horizontal\ center\ of\ mark\ bounding\ box$ *Note: For left-right diagonal, the diagonal region was bounded by two lines: - The first line's start coordinate: (¼ width of bounding box, 0) and end coordinate: (width, ¾ height of bounding box) - The second line's start coordinate: (0, ¼ height of bounding box) and end coordinate: (¾ width of bounding box, height). - A similar process was used for right-left diagonals. Checks and crosses have distinct regions of darker pixels, which can be captured by the diagonal, horizontal and vertical regions. Checks tend to have one diagonal with more dark pixels, while crosses have both diagonals. Crosses have a distinct mark in the horizontal region of the mark bounding box if the region captures the intersection of the diagonals in the cross.

Lightly Filled	$\sigma_{orig} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2}$ Lightly filled marks have a low standard deviation of pixel intensities inside the ellipse because these marks fill the entire mark target evenly. Other marks like checks and crosses have a high contrast of high and low pixel intensities that don't fill the entire mark target.
Overfilled	$v_{orig} = \mu_{intensityInsideTarget}$ $v_{orig} = \mu_{intensityOutsideTarget}$ $v_{orig} = \mu_{intensity}$ Overfilled marks, on average, have more dark pixels inside and outside the ellipse compared to other marginal marks.

Visual examples of each of the nine customized mark features show how the custom features capture the salient characteristics of each mark type (see Figure 7). The objective of designing customized features is to improve characterization of a particular mark type by minimizing intra-class separation. The second objective is to improve discrimination by maximizing inter-class separation.

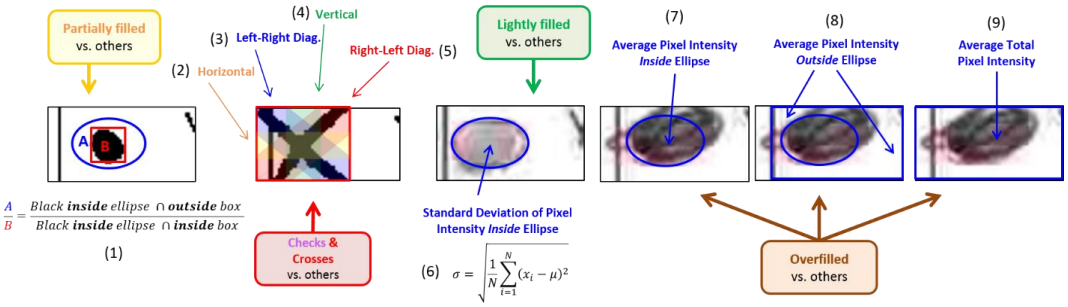

Figure 7. Design of mark image feature with focus on better characterization (intra-class) and discrimination (inter-class)

The parallel coordinates chart (see Figure 8) is used to visualize the capability of the nine custom features in discerning the mark types. Each of the five mark types corresponds to a unique color along with the feature vector values for each mark image. Figure 8 demonstrates the expected separation between mark types at each feature.

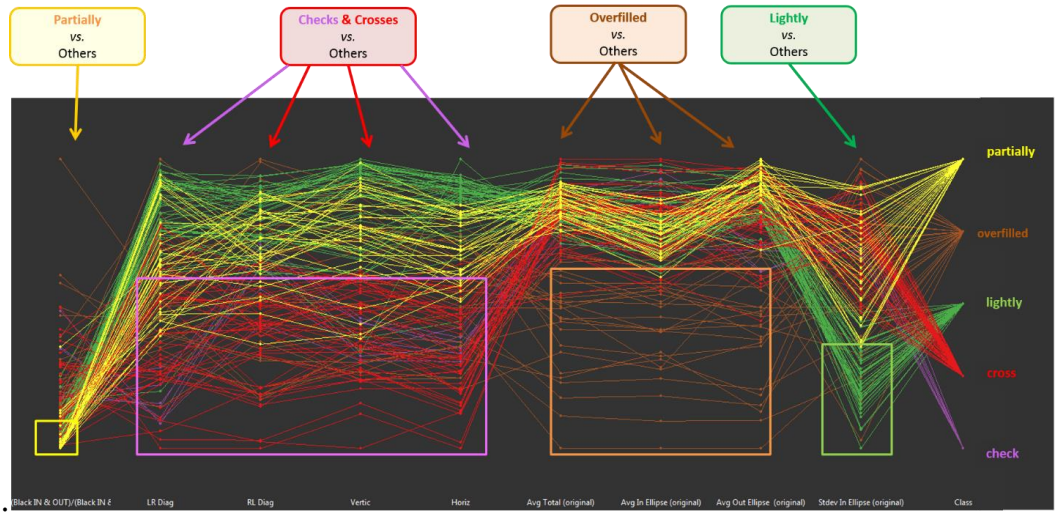

Figure 8. Parallel coordinates plot for 9 custom features

In addition to qualitative inspection using parallel coordinates charts, these nine custom features were evaluated by analyzing their performance on synthetic mark images with known mark types. Synthentic test samples were generated with the intention to supplement the analysis with more training and test data to provide more rigorous experimental validation of the nine custom features. We generated synthetic marks with variations in location (with respect to mark target), line width, average pixel intensity, and vertical stretch (see Figure 9). Synthetic cross marks and check marks had variations in location offset from mark target center (*cross:* [-10, 10], Δ=2, *check:* [-10, 8], Δ=2), vertical stretch (*cross:* [0, 5], Δ=1, *check:* [0, 6], Δ=2), and line width (*cross:* [2, 5], Δ=1, *check:* [2, 5], Δ=1). Lightly filled and overfilled marks had one degree of freedom with lightly filled varying in intensity from 127 to 255 with Δ=10 and overfilled varying in the overlap of strokes and mark target from 50 % to 100 % overlap and Δ=10. Partially filled marks varied in their location (random pixel amount [0, 8]) and size of mark (rounded rectangle parameterized by: $width \cdot height = (6 + 2i)(4 + i), 0 \leq i \leq 10, \Delta i = 2$). All variations are in pixels. Note that the synthetic models are based on our observations since there is no standard definition of marginal mark types.

Figure 9. Sample synthetic mark images for checks, crosses, lightly filled, overfilled, and partially filled

Given the various synthetic marks and the training mark data, we compared parallel coordinate charts for the four features used to distinguish checks and crosses from partially, overfilled, and lightly filled marks as shown in Figure 10. We observed similar trends amongst the features with both the synthetic and training data in Figure 10 left. However, we note that the synthetic marks did not capture the amount of variation present in the training data. This can be seen in Figure 10 right through the cross data (red cluster) in terms of its location and spread. Thus, evaluation based on synthetic data requires a wider range of parameters in order to properly capture the variation of the marginal marks extracted from real ballots. For example, in actual marginal marks, check marks and crosses may also vary by intensity, size and even shape. Further visual analysis and experimentation in generating the synthetic marks is needed to define the variables for actual marginal marks.

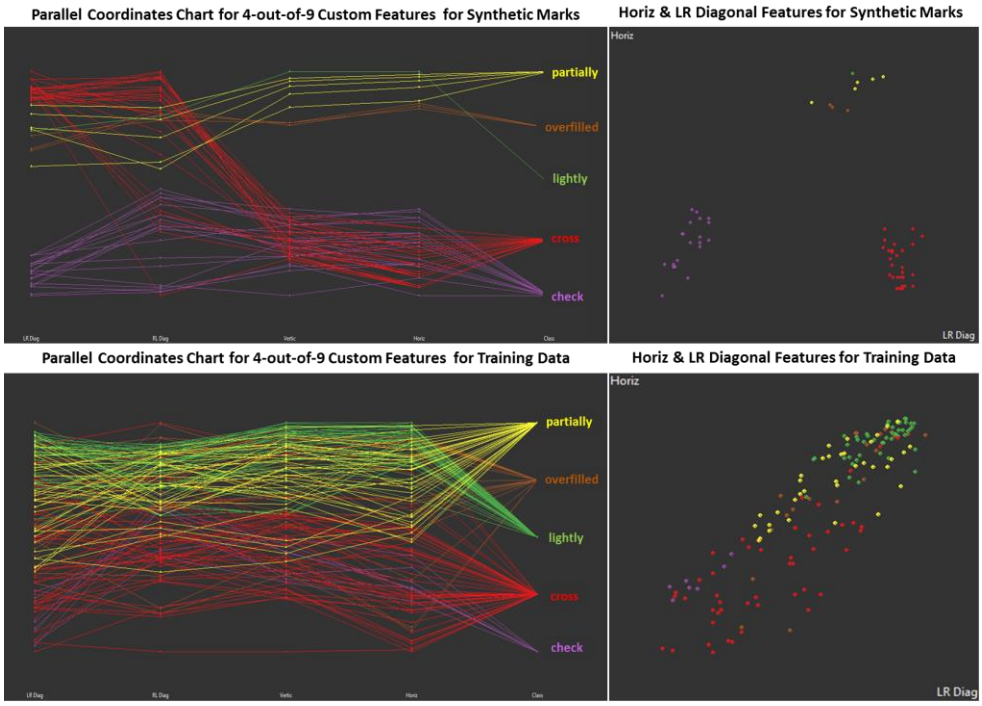

Figure 10. Comparison of custom features (four features for checks & crosses are pictured) between synthetic marks and marks from the training dataset.

4. Classification Accuracy Evaluation

In order to assess the mark-type classification accuracy based on our three features sets, the three feature value sets were imported into the WEKA[1] machine learning software [14] for analysis. We evaluated the J48 Tree Classifier, Simple Logistic Regression Model, Logistic Regression Model, and Multilayer Perceptron over all 7 combinations of the three sets of features as shown in Table 2. While decision trees have the advantage of establishing clear rules, it is limited by its ability to handle more ambiguous cases. Non-decision tree models such as Simple Logistic, Logistic, and Multilayer Perceptron provide a probabilistic model, meaning they provide a probability that a certain mark belongs to a specific category. For fusion cases, the probabilities may be more evenly distributed between two mark types.

Each classifier used 10-fold cross validation and was executed on the feature value file for each of the three sets of features (33 Fiji features, 22 baseline features, and 9 custom features) and for subsequent combinations of features. ***Table 3*** summarizes the 10-fold cross validation learning algorithm evaluations for the 168 labeled marginal mark images and their accompanying image feature sets. Based on ***Table 3***, the best combination of feature sets is with 9 custom or 9 custom plus 22 baseline features. Simple Logistic and Multilayer Perceptron achieved the highest classification with our custom feature set.

Although Simple Logistic and Multiple Perceptron models achieved the highest classification accuracy with 9 custom features, we also selected the J48 Tree Classifier for modeling. The J48 Tree Classifier is WEKA's version of C4.5 decision tree algorithm, where attribute splits are determined based on information entropy [15]. The ability to create clear rules in classifying discrete classes is one of the main advantages of decision trees. We note that the J48 Tree Classifier achieves the highest classification accuracy of 90.4762 % when using the feature set with 9 custom + 22 baseline features. However, we anticipated poorer performance of the J48 Tree Classifier as compared to the non-decision tree classifiers due to its inability to properly handle more ambiguous marks with feature values that fall in a continuum between two mark types. Although more complex to interpret, we expected the non-decision tree models to better account for fusion mark types.

Table 3. Classification accuracy comparison for seven unique combinations of image feature sets

Classifier / Feature Sets	J48 Decision Tree	Simple Logistic	Logistic	Multilayer Perceptron
Fiji (33 features)	86.4198 %	88.8889 %	85.8025 %	89.5062 %
22 baseline	86.9048 %	89.2857 %	86.3095 %	92.2619 %
9 custom	85.119 %	**94.0476 %**	89.881 %	**94.6429 %**
Fiji + 22 baseline	86.4198 %	90.1235 %	83.9506 %	88.2716 %
Fiji + 9 custom	84.5679 %	90.1235 %	83.9506 %	88.8889 %
9 custom + 22 base	**90.4762 %**	91.6667 %	**92.2619 %**	92.8571 %
All (64 features)	85.1852 %	90.1235 %	83.9506 %	87.6543 %
Delta (Best - Fiji)	4.0564 %	5.1587 %	6.4594 %	5.1367 %

In terms of developing a model to probabilistically classify a marginal mark type, the Simple Logistic Regression provides a good tradeoff between ease of interpretation and reasonable accuracy at 94 percent with 10-fold cross validation. Utilizing WEKA and the Simple Logistic Regression [16], we generated a probabilistic model based on the nine normalized custom features. In simple logistic regression, each model is binary. For example, the model can classify whether the marginal mark belongs to a single class or not. The logistic regression model, Equation 3, fits the log odds with linear function through the custom set of features. For each mark type, M_i, we have the probability function, Equation 5, derived from the odds function, Equation 4, where x_k is the kth feature, Θ_k is the optimized regression coefficient for the predictor x_k and X is the vector of features $[x_1,\ldots,x_9]$. The regression coefficients are optimized iteratively over the training set until the cost function converges to a minimum. The cost function is the square of the difference between the predicted class and the observed class. ***Table 4*** shows the coefficients for each feature, which can be interpreted as the weight of the feature. Simple logistic regression not only provides a probability model for each marginal mark type, it also enables us to see the significance of each feature with respect to a specific marginal mark type.

$$logit(\pi_{M_i j}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 \ldots + \theta_9 x_9 \quad (3)$$

$$\pi_{M_i j} = \exp(\, logit(\pi_{M_i j})) \quad (4)$$

$$P(M_i = 1 | X) = \frac{\pi_{M_i j}}{1+\pi_{M_i j}} \quad (5)$$

Table 4. *Simple logistic model coefficients for the 5 mark types and 9 normalized custom attributes.*

x_k / M_i	Pixel Ratio	Horiz. Pixels	Left-Right Diagonal	Right-Left Diagonal	Vertical	$\sigma_{pixelsInside}$	$\mu_{intensityInside}$	$\mu_{intensityOutside}$	$\mu_{intensity}$
Check mark	0	0	-31.6	21.31	-2.7	42.91	55.03	0	0
Cross	-6.64	-28.71	3.62	-4.66	-5.42	24.86	0	2.56	3.3
Lightly filled	4.03	47.27	0	0	0	-40.08	0	-24.31	-2.72
Overfilled	9.48	0	0	0	4.23	-19.14	-44.98	-3.72	-33.49
Partially filled	0	18	-11.26	-1.75	13.05	0	0	17.17	12.53

Multinomial logistic regression [17] generalizes the simple logistic regression to address the multiclass problem. Similar to simple logistic regression, the coefficients are determined iteratively by minimizing the cost function. The optimization is based on maximum log-likelihood, which is prone to overfitting the training set. The problem is exacerbated when there are large number of features and a relatively small training set. Table 3 shows how the logistic regression had reduced classification accuracy, most likely due to overfitting of the training set.

Multilayer perceptron (MLP) is a type of neural network model which can be applied to noisy data to extract statistically significant patterns. The accuracy of the MLP is among the highest, but with only a difference of 0.6 % from simple logistic regression. To ensure ease of model interpretability, the probabilistic models derived from the simple logistic regression appear sufficient.

5. DISCUSSION

The marginal mark classification should be compared to the state of the art methods that use primarily pixel intensity features and other proprietary features. Since these methods are proprietary, we could not quantify the improvement of our custom features compared to existing mark detection algorithms deployed in optical mark scanners used for voting.

By applying 4 statistical classifier models (both tree and non-decision tree models) to the extracted marginal marks, we were able to classify mark types with accuracy greater than 90% and demonstrated improvement in mark-type classification using our custom feature set when compared to off-the-shelf features between 4.1 % and 6.5 %. We achieved the classification improvement by incorporating knowledge of handwritten marginal mark patterns into custom mark image features.

Analyzing the results of the four classifiers, the simple logistic regression models provided the optimal combination of interpretability and accuracy with the custom features. The binary probabilistic classification models are presented for each class type. The models demonstrate the feasibility of defining marginal mark types based on a linear combination of weighted features.

6. CONCLUSION & FUTURE WORK

The work presented is intended to demonstrate the feasibility of using custom features for automated classification of marginal marks. Five semantic labels for marginal marks deemed to be representative of given ballots were defined. We designed 9 custom feature extractors for each of the five marginal mark types for both binary and color images. A rigorous experimental validation and uncertainty analysis of the custom features would determine whether the process of mark characterization can be used to serve as a means for developing quantitative thresholds for testing and calibrating optical scan voting equipment. The quantitatively defined parameters enable manufacturers to develop repeatable and unbiased means of testing whether the equipment meets specification as discussed in the draft of VVSG 1.1. The quantitative features can also enable states to establish quantitative specifications in the definition of marginal marks.

A systematic processing pipeline with customized features resulting in quantitative probabilistic models for classifying marginal marks has also been demonstrated. This experimental methodology should ultimately be applied onto the entire dataset of ballot images as well as to ballots with different assumptions. For example, ballots might not have the assumed attributes in general due to state-to-state variations of ballot templates. For example, aside from ellipse mark targets, other ballots have broken arrow or square targets, each with a unique set of marginal mark types. Furthermore, ballot level features (such as the consistency and pattern of a user's vote mark across a single ballot) could be analyzed for better understanding of mark type patterns. Future research of additional mark types to those explored in this paper would provide a quantitative characterization to clearly and consistently classify "fusion" cases into a distinct category where human assessors may have difficulty reaching consensus. In terms of the analysis presented in this work, ideally more than 10 human assessors would be desirable to increase the statistical confidence in the ground truth. The method of generation synthetic marks was also designed and implemented

to supplement the actual data. From the analysis additional variables need to be introduced to increase the fidelity of the synthetic marginal marks to the types of marginal marks that exist. The ability to generate synthetic marks with high fidelity can further improve testing of optical scanners in handling marginal marks.

The ability to define mark types with quantitative features and classify mark types using a probabilistic model serves as an initial step to enabling the optical scan manufacturers to ensure consistency and transparency in the interpretation of a wide range of manual marks as votes or non-votes. It is only through quantitative characterization and classification that enables objective definitions and decisions to be made when ambiguity arises in manually marked ballots.

7. REFERENCES

[1] T. Antonyan, S. Davtyan, S. Kentros, A. Kiayias, L. Michel, N. Nicolaou, A. Russell and A. Shvartsman, "State-wide elections, optical scan voting system and the pursuit of integrity," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, 2009.

[2] ProCon.org, "Voting Systems & Use: 1980-2012," ProCon.org, 6 February 2013. [Online]. Available: http://votingmachines.procon.org/view.resource.php?resourceID=000274. [Accessed 20 November 2014].

[3] D. Lopresti, G. Nagy and E. Smith, "A document analysis system for supporting electronic voting research.," in *The Eighth IAPR International Workshop on Document Analysis Systems '08*, Nara, 2008.

[4] D. Jones, "On optical mark sense scanning," in *Towards Trustworthy Elections*, Springer, 2010, pp. 175-190.

[5] U.S. Election Assistance Commission, "2012 Draft Voluntary Voting System Guidelines Version 1.1," 31 August 2012. [Online]. Available: http://www.eac.gov/assets/1/Documents/VVSG%20Version%201.1%20Volume%201%20Public%20Comment%20Version-8.31.2012.pdf. [Accessed 4 March 2015].

[6] D. Flater, "NIST and the Help America Vote Act (HAVA)," 1 December 2006. [Online]. Available: http://www.nist.gov/itl/vote/upload/Marginal.pdf. [Accessed June 2014].

[7] D. Lopresti, G. Nagy and E. Smith, "Document analysis issues in reading optical scan ballots.," in *DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*, New York, 2010.

[8] A. Cordero, T. Ji, A. Tsai, K. Mowery and D. Wagner, "EVT/WOTE'10," 8 September 2010. [Online]. Available: https://www.usenix.org/legacy/event/evtwote10/tech/full_papers/Cordero.pdf. [Accessed July 2014].

[9] T. Ji, E. Kim, R. Srikantan, A. Tsai, A. Cordero and D. Wagner, "USENIX EVT/WOTE'11," 10 August 2011. [Online]. Available: https://www.usenix.org/legacy/event/evtwote11/tech/final_files/Ji.pdf. [Accessed July 2014].

[10] D. W. Jones, "Counting Mark-Sense Ballots," February 2002. [Online]. Available: http://homepage.cs.uiowa.edu/~jones/voting/optical/. [Accessed February 2015].

[11] Michigan Election Law, "Michigan Compiled Laws Complete Through PA 572 of 2014 & 1 & 2 of of 2015," 5 March 2015. [Online]. Available: https://www.legislature.mi.gov/%28S%28mjsnu3avejk00s55aoymxe45%29%29/documents/mcl/pdf/mcl-168-803.pdf. [Accessed 5 March 2015].

[12] "Humboldt County Election Transparency Project," July 2009. [Online]. Available: http://humtp.com/. [Accessed June 2014].

[13] F. Zhao, Q. Huang and W. Gao, "Image Matching by Normalized Cross-Correlation," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, Volume: 2*, Toulouse, 2006.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, *The WEKA Data Mining Software: An Update,* vol. 11, SIGKDD Explorations, 2009.

[15] R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[16] M. Summer, E. Frank and M. Hall, "Speeding up logistic model tree induction," *9th European Conerence on Principles and Practice of Knowledge Discovery in Databases,* pp. 675-683, 2005.

[17] S. le Cessie and J. van Houwelingen, "Ridge Estimators in Logistic," 1992.