Towards The Collection of Census Data From Satellite Imagery Using Data Mining: A Study With Respect to the Ethiopian Hinterland

Kwankamon Dittakan, Frans Coenen and Rob Christley

Abstract The collection of census data is an important task with respect to providing support for decision makers. However, the collection of census data is also resource intensive. This is especially the case in areas which feature poor communication and transport networks. In this paper a method is proposed for collecting census data by applying classification techniques to relevant satellite imagery. The test site for the work is a collection of villages lying some 300km to the northwest of Addis Ababa in Ethiopia. The idea is to build a classifier that can label households according to "family" size. To this end training data has been obtained, by collecting on ground census data and aligning this up with satellite data. The fundamental idea is to segment satellite images so as to obtain the satellite pixels describing individual households and representing these segmentations using a histogram representation. By pairing each histogram represented household with collated census data, namely family size, a classifier can be constructed to predict household sizes according to the nature of the histograms. This classifier can then be used to provide a quick and easy mechanism for the approximate collection of census data that does not require significant resource.

Keywords: Data Mining, Image Classification, Satellite Image Analysis, Satellite Image Mining, Census Analysis, Census Mining

---

Kwankamon Dittakan
Department of Computer Science, University of Liverpool, Ashton Building , Aston Street, L69 3BX Liverpool, UK e-mail: dittakan@liverpool.ac.uk

Frans Coenen
Department of Computer Science, University of Liverpool, Ashton Building , Aston Street, L69 3BX Liverpool, UK e-mail: coenen@liverpool.ac.uk

Rob Christley
Institute of Infection and Global Health, University of Liverpool, Leahurst Campus, Chester High Road, CH64 7TE Neston, Cheshire,UK e-mail: robc@liverpool.ac.uk

1 Introduction

A census is a collection of information about the nature of a population of a given area. Census collection tends to be undertaken using a questionnaire format. Questionnaires are usually either distributed by post or electronically for self-completion, or completed by field staff. There are many problems associated with the collection of census data, especially in the case of national censuses. The first problem is census budget, the collection of census data requires a considerable resource in terms of money and “manpower”. Another problem is the cost of processing the data after it has been collected. A third issue is that there is often a lack of good will on behalf of a population to participate in a census, even if they are legally required to do so, because people are often suspicious of the motivation behind censuses (especially when collected by government organisations) [7]. These problems are compounded in areas where there are poor communication and transport infrastructures; and/or an extensive, but sparsely populated, hinterlands.

The solution proposed in this paper is founded on the idea of constructing classifiers that can predict census information according to the nature of satellite views of households. In some areas, such as inner city areas, this is unlikely to be appropriate; however in sparsely populated rural areas, as will be demonstrated, this can provide a effective and efficient mechanism for collecting census data. The fundamental idea of the proposed research is that, given a set of training data describing households and their geographical locations, we can obtain satellite images of these households and use image classification techniques to construct a classifier that can be applied to the entire region and consequently automatically collect census data at very low cost. Of course this will be an approximation, there is always a trade of between resource reduction and accuracy; and, as already noted, is likely to be more effective in rural and suburban areas, than in city centres and commercial areas.

The rest of this paper is organised as follows. In Section 2 some previous work is presented. Section 3 then provides detail of the proposed census mining framework, including reviews of the image enhancement, image segmentation, feature representation and feature selection mechanisms adopted. Section 4 reports on the evaluation of the framework (conducted using a sequence of experiments applied to data collected from villages lying some 300km to the northwest of Addis Ababa in Ethiopia). Finally, a summary and some conclusions are presented in Section 5.

2 Previous work

Satellite image interpretation offers advantages with respect to many applications. Examples include: land usage, regional planning, forest area monitoring, forest mapping, disease and fire detection and wild life studies [4, 13, 11]. The satellite image interpretation technique proposed in this paper is founded on image classification, the process of applying classification techniques to image sets. A typical generic image classification task is to label image data according to a predefined set of classes. In the context of satellite imagery; image classification has been used, for example, for linking urban land cover (obtained from satellite imagery) with urban function characteristics obtained from population census data [14]. However, there is little reported work on the application of image classification techniques to satellite image data for the purpose of census data collection although one example can be found in [19] where classification techniques are applied to satellite images to obtain information such as number of households, total population and urban population at the sub-district level. The distinction between this work and that described in this paper is that the proposed census mining framework operates at a much finer level of granularity.

Image classification, and by extension satellite image classification, requires images to be represented in some form that lends itself to the application of image classification techniques. There are many techniques that have been proposed for representing image features. In general these can be separated into three categories: (i) text based, (ii) semantic based and (iii) content based [10]. In the first category images are represented simply using keywords which describe each image. In the second category researchers have tried to capture the semantic meaning of images. In the third category some general content such as colour, texture and shape, is used to represent entire images or parts of images. The advantage of the third is that this representation can be automatically generated, and thus this is the technique adopted with respect to the work described in this paper where the proposed representation is founded on image colour.

One method of encapsulating image colour is to represent the distribution of colours using histograms [6, 12, 20]. This can be applied to entire images or parts there of. In the case of the satellite imagery of interest we wish to capture the parts of images that represent households. This in turn requires some segmentation process. Image segmentation is the process of partitioning an image so as to identify objects within the given image. Popular segmentation techniques include Threshold segmentation [16, 17], Region-growing segmentation [2, 18], Edge-based segmentation [15] and Texture-Based segmentation [24]. Because the households of interest tended to be rectangular in shape, the Edge-based segmentation technique was adopted for the purpose of identifying (segmenting) households.

3 Census Mining Framework

The proposed census framework is described in this section, the framework comprises six stages: (i) coarse segmentation, (ii) image enhancement, (iii) detailed segmentation, (iv) representation, (v) feature section and (vi) classifier generation. The proposed census mining framework takes as input a set of labelled satellite images, feeds then into an appropriately trained classifier and outputs prescribed census data. The nature of the classifier will depend on the nature of the desired census data. Different types of census data will require different types of classifier. The focus of the work describing in this papers is household size in terms of number of people normally resident in a given household. To build such a classifier suitable labeled training data was required. To act as a focus for the work the research team arranged for the manual collection of census data in a rural area lying some 300 km to the northwest of Addis Ababa in Ethiopia (Figure 1), which could then be used to generate an appropriate training data set.

Fig. 1 The test site location: Harro district in Ethiopia (indicated by arrow)

The collected census data included geographical coordinates so that individual households could be related to satellite imagery. The research team used Google Earth imagery but clearly other forms of satellite imagery could equally well have been used. The identified households could therefore be located and segmented so that a collection of household images could be obtained. The first stage in the framework was thus the coarse segmentation of satellite imagery to isolate individual households or groups of households so that a collection of household sub-images was arrived at. A typical satellite image is presented in Figure 2, satellite images of this forum will thus first be roughly segmented to give N household sub-images. Note that in some cases the sub-images may still contain several households.

The next stage in the process is image enhancement; this is briefly described in Sub-section 3.1. Once suitable enhancement had been applied the third stage was the detailed segmentation of the household sub-images so that individual households could be identified. The proposed segmentation process is described in Sub-section 3.2. The fourth stage was to represent the images in such a way that some image classification technique could be applied. As already noted in Section 2, there are a great many image representation techniques that have been proposed in the literature. The technique used for the proposed census mining was founded on a histogram based representation from which various statistics could be extracted. The acquired histogram data could then be coupled with class data (family size with

Fig. 2 Example satellite image showing individual households (Harro district in Ethiopia)

respect to the evaluation presented in this paper) and used to generate the desired classifier. The process for generating histograms is presented in Sub-section 3.3. It is commonly accepted that not all features are significant with respect to individual image classification tasks, and that it is desirable (for computational resource saving reasons) to limit the dimensionality of the feature space, hence a feature selection process was also applied. This is presented in Sub-section 3.4. Once a suitable set of features had been identified, expressed in terms of a feature vector, a suitable classifier generator could be applied (stage 6) to produce the final desired classifier.

3.1 Image Enhancement

This sub-section describes the image enhancement processes applied to the input data. An example household sub-image, obtained from a satellite image of the form shown in Figure 2, is presented in Figure 3(a). However, before image enhancement can be applied it was first necessary to register and align the images so that each household (delimited by its surrounding rectangular boundary) was aligned in a north-south direction. The purpose of this registration and alignment was to facilitate future segmentation. The result is a shown in Figure 3(b), where the image shown in Figure 3(a) has been appropriately aligned.

The main issue to be addressed during the image enhancement is that the colours within the collected sub-images are frequently not consistent across the image set. In our work histogram equalisation [11, 9, 8] was applied to the satellite image data

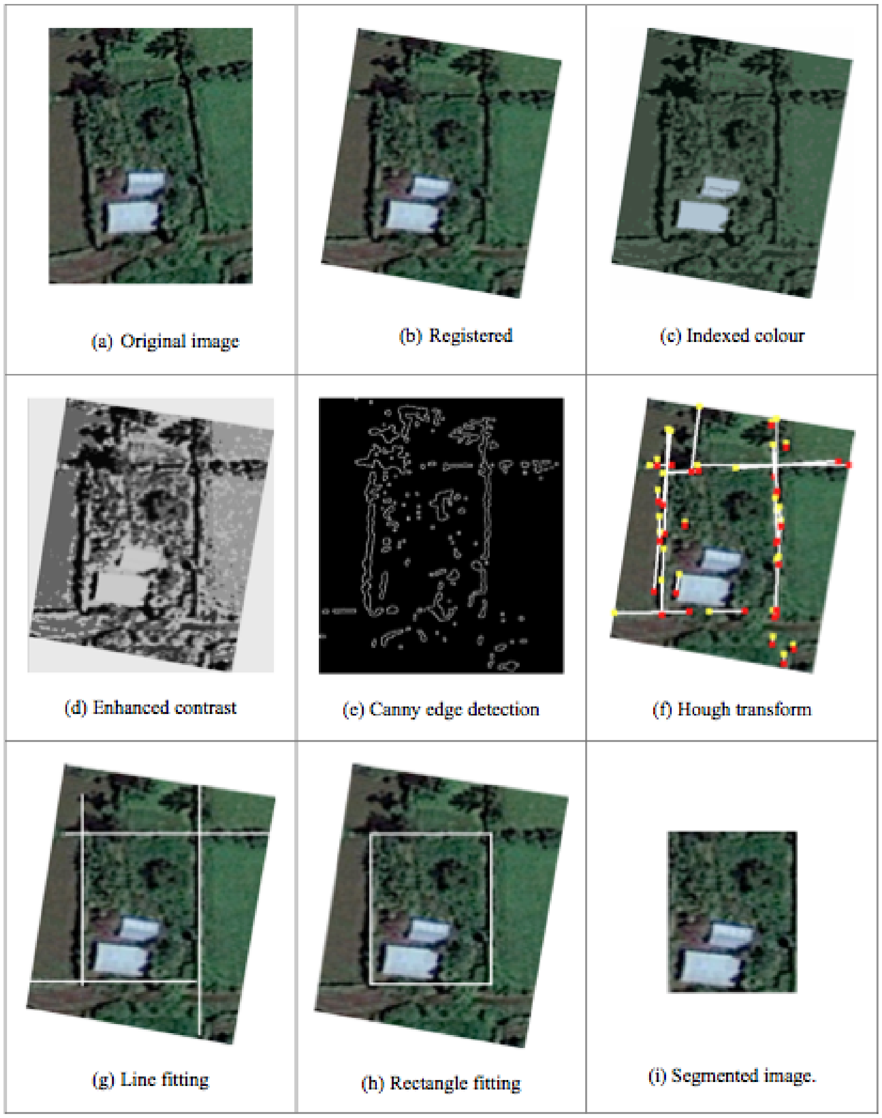

Fig. 3 Sub-image processing

so as to obtain a consistent colour regime. The first step was to convert each RGB colour represented images into an eight colour indexed image (as shown in Figure 3(c)), and then transform them to grayscale images to which histogram equalisation was then applied. Histogram equalisation is a method for image enhancement directed at ensuring an equal colour distribution (Figure 3(d)). The process commences by selecting a reference image which is then used for normalisation purposes with respect to the remaining images.

3.2 Image Segmentation

Once the image enhancement process was complete the next stage was to segment the images so as to isolate individual households. It has already been noted that the households of interest are typically defined by a rough rectangular boundary in which buildings and related objects are located. We wished to segment these images so that these rectangular areas can be clearly isolated, however the boundaries are frequently not well defined in that the edges are not continuous. Thus, for example, region-growing segmentation techniques would be unlikely to perform well. The Canny edge detection algorithm [3] and the Hough Transform [5] were thus applied. Canny edge detection is an edge detection technique for identifying object contours from intensity discontinuities. The Hough transform is a segmentation technique suited for identifying imperfect instances of objects of certain predefined shapes (such as a straight lines). Prior to applying Canny edge detection and the Hough transform, contrast adjustment was applied to the images so that the household boundaries could be more readily distinguished. Canny edge detection was then applied, the result is shown in Figure 3(e) were the detected edges have been highlighted. As a result of applying the Hough transform, we have a collection of "lines" as shown in Figure 3(f). Each line is defined by a start and end point, and a ρ and θ value (length and direction).

We now need to fit a rectangle to this set of lines. This was achieved by applying a Least squares approach [1] applied to each group of lines approximating to the top, bottom, left and right sides of the rectangle. As results the rectangle surrounding each household was demarcated by a pair of horizontal and a pair of vertical lines (Figure 3(g)). The intersections of the lines can then be found so as to delimit the surrounding rectangle in term of its four corners (Figure 3(h)). The final result is a set of segmented household images as shown in Figure 3(i).

3.3 Features Representation

After the households have been segmented each had to be represented using some appropriate mechanisms that serves to: (i) capture the salient features of the object in a concise manner and (ii) permit the ready application of some classification algorithm. As already noted above, a histogram based representation is proposed. Four histograms were generated for each household describing the three RGB colour channels and a grayscale version. Once the histograms had been generated (one per household) two categories of feature could be extracted: histogram features and statistical features. For the histogram features, each of the four histograms was divided into 16 “bins”, giving 64 features in total. Four categories of statistical feature was considered: (i) general features (G), (ii) entropy features (E), (iii) grey-level co-occurrence matrix features (M) [23] and (iv) wavelet transform features (W) [22]. The complete set of statistical features (19 in total) is presented in Table 1, the letter in parenthesis behind each feature type indicates that category of the feature.

Table 1 Statistical features.

#	Description	#	Description	#	Description
1	Average red (G)	9	Standard deviation (G)	16	Average approximation
2	Average green (G)	10	Entropy (E)		coefficient matrix, *cA* (W)
3	Average blue (G)	11	Average Local Entropy (E)	17	Average horizontal
4	Average hue (G)	12	Contrast (M)		coefficient matrix, *cH* (W)
5	Average saturation (G)	13	Correlation (M)	18	Average vertical
6	Average value (G)	14	Energy (M)		coefficient matrix, *cV* (W)
7	Average grayscale (G)	15	Homogeneity (M)	19	Average diagonal
8	Mean (G)				coefficient matrix, *cD* (W)

3.4 Feature Selection

The 64 histogram and 19 statistical features identified (as described above) were used to describe each household. Prior to classifier generation, so as to reduce the number of features to be considered to those that best served to distinguish between different classes of household, a feature selection process was applied. Five feature selection strategies were considered: (i) χ-squared, (ii) Gain Ratio, (iii) Information Gain, (iv) One-R and (v) Relief-F. Note that the χ-squared measure evaluates the “worth” of attributes by computing and comparing their χ-squared value. The Gain ratio, as the name suggests, evaluates attributes by measuring gain ratio. Information gain evaluates the attributes by evaluating the entropy of the class, which characterises the purity of an arbitrary. The One-R (One Rule) feature selection strategy comprises a simple classifier that generate a one-level classifier. The Relief-F method is an instance-based technique, which randomises the instances and checks the neighboring instances (records) for the same and different class labels.

4 Evaluation

To evaluate the proposed process, as already noted, labeled training data was collected from a rural region 300km to the northwest of Addis Ababa in Ethiopia. More specifically data was collected from the Horro district in the Oromia Region of Ethiopia, an area bounded by the 9.5N and 9.8N parallels of latitude and the

37.0E and 37.5E lines of longitude as shown in Figure 1. Data, including family size, latitude and longitude for each household was collected by University of Liverpool field staff in May 2011. The minimum and maximum family size was 2 and 10 respectively, the mean was 5.97 and the medium 6. Thus, for evaluation purposes the households were divided into two classes: less than 7 people ("Small family") and greater or equal to 7 ("Large family"). The required household satellite image data was extracted from GeoEye satellite images which had a 50 centimetres ground resolution in Google Earth. The high resolution of these images made them well suited with respect to the desired census mining because objects such as buildings and fields could easily be identified. Figure 2 presented a sample satellite image covering part of the Harro district. From the Figure several households can clearly be observed. In total, data for 30 households was processed for evaluation purposes.

An overview of the experimental set up is given by the schematic shown in Figure 4. The data was processed as described in Section 3 above. Data discretisation was then applied to the selected features so that each continuously valued attribute was converted into a set of ranged attributes. Then the classification learning methods were applied. The Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench [21] was used for classifier generation purposes. 10-fold cross-validation was used throughout. The performances of each classifier was recorded in terms of: (i) accuracy, (ii) sensitivity, (iii) specificity, (iv) precision with respect to the small family, (v) precision with respect to the large family, (vi) F-measure and the (vii) ROC area[1].

Extensive evaluation has been conducted with respect to the proposed techniques. This section reports on only most significant results obtained (there is insufficient space to allow for the presentation of all the results obtained). The evaluation presented in this paper were directed at four goals. The first was to determine the effect of classification performance using either histogram features only, statistical features only and a combination of the two (Sub-section 4.1). The second compared the operation of the various suggested feature selection algorithms (Sub-section 4.2). The third was directed at an analysis of the effect that the number of selected attributes, K, had on performance (Sub-section 4.3). Finally, the fourth was directed at determining the effect on classification performance as a result or using different learning methods (Sub-section 4.4).

4.1 Data Representation (Histogram Attributes v. Statistical attributes)

Three different data sets were generated, to conduct the experiments described in this paper, in order to investigate the effect of different types of data on classification performance. The three data set were as follow: (i) Histogram, comprised

[1] For calculation of the evaluation metrics a confusion matrix with "small family" as the positive class was used.

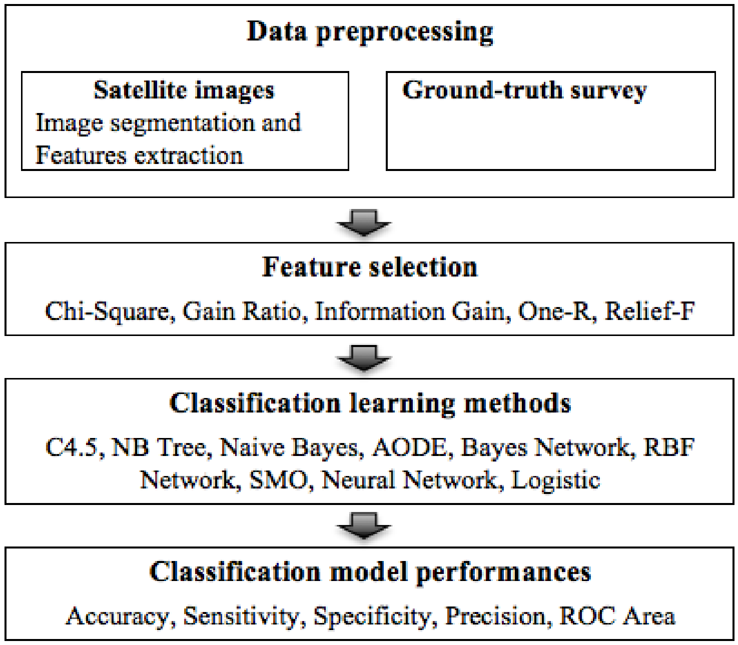

Fig. 4 Schematic illustration the evaluation set up

of the 64 histogram bin attributes that may be extracted from the four different identified histogram representations; (ii) Statistical, comprising the 19 attributes (including general features, entropy features, grey-level co-occurrence matrix features and wavelet transform measures); and (iii) Combined, which combined the two and hence comprised 83 (64+19) attributes. For the experiments information gain was used as the feature selection method together with $K = 15$ and a Bayesian Network learning method because (as will be seen from the following reported results) it was found these tended to generated the best performance. The value of $K = 15$ was selected because initial experimentation (not report here) indicated that this was the most appropriate. Table 2 shows the classification results produced using the three data sets. Form the table it can be seen that the Combined data produced the best overall classification performance, probably because of the greater quantity of data (number of attributes) used to generate the classifier in this case. It is worth noting that very good results were produced using the Combined data set.

4.2 Feature selection

Recall that five kinds of feature selection techniques were considered: (i) χ-squared (ChiSquaredAttributeEval), (ii) Gain ratio (GainRatioAttributeEval), (iii) Informa-

Table 2 Classification performance using data sets comprising different kinds of feature ($K = 15$, Information gain feature selection and Bayesian Network learning).

Data Set	Accu-racy	Sensi-tivity	Speci-ficity	Precision "Small"	Precision "Large"	F-Measure	ROC Area
Histogram	0.867	0.882	0.846	0.882	0.846	0.882	0.923
Statistical	0.600	0.706	0.462	0.632	0.545	0.667	0.688
Combined	0.933	1.000	0.846	0.895	1.000	0.944	0.959

tion gain (InfoGainAttributeEval), (iv) One-R (OneRAttributeEval) and (v) Relief-F (ReliefFAttributeEval). For the experiments used to compare these five methods the Combined data set was used as this had been found to produce the best results as established in the previous sub-section. $K = 15$ was again used together with a Bayesian Network learning method forth the same reasons as before (because they had been found to produce the best results). The results from the experiments are present in Table 3. From the table it can be seen that the χ-square and Information Gain feature selection techniques produced the best results; although, in the context of the ROC area value, information gain outperformed the χ-squared measure, and thus it can be concluded that information gain is the most appropriate measure in the context of the census mining application considered in this paper.

Table 3 Classification performance using different feature selection techniques (Combined, $K = 15$ and Bayesian Network learning).

Selection techniques	Accu-racy	Sensi-tivity	Speci-ficity	Precision "Small"	Precision "Large"	F-Measure	ROC Area
χ-Square	0.933	1.000	0.846	0.895	1.000	0.944	0.928
Gain Ratio	0.900	1.000	0.769	0.850	1.000	0.919	0.950
Information Gain	0.933	1.000	0.846	0.895	1.000	0.944	0.959
One-R	0.800	0.941	0.615	0.762	0.889	0.842	0.851
Relief-F	0.733	0.941	0.462	0.696	0.857	0.800	0.928

4.3 Numbers of attribute

In order to investigated the effect on classification performance of the value of K a sequence of experiments was conducted using a rang of values of K from 5 to 25 incrementing in steps of 5. For the experiments the Combined data and information gain feature selection were used, because they had already been shown to produce good performance, together with Bayesian Network learning. Table 4 shows the classification results produced. From the table it can be seen that $K = 10$ and $K = 15$ produced the best performance, although, as in the case of the experiments conducted to find the most appropriate feature selection mechanism, when the ROC area value was considered $K = 15$ produced a slightly better performance than when $K = 10$ was used. Thus it can be concluded that $K = 15$ is the most appropriate measure in the context of census mining, as considered in this paper. It is conjectured that low values of K did not provide a good performance because there was insufficient data to built an effective classifier, while large values of K resulted in overfitting.

Table 4 The classification performances using different values for K (Combined, Information Gain and Bayesian Network learning).

Number of attributes	Accuracy	Sensitivity	Specificity	Precision "Small"	Precision "Large"	F-Measure	ROC Area
5 attributes	0.900	1.000	0.769	0.850	1.000	0.919	0.946
10 attributes	0.933	1.000	0.846	0.895	1.000	0.944	0.950
15 attributes	0.933	1.000	0.846	0.895	1.000	0.944	0.959
20 attributes	0.900	0.941	0.846	0.889	0.917	0.914	0.964
25 attributes	0.900	0.941	0.846	0.889	0.917	0.914	0.946

4.4 Learning methods

Nine learning methods were considered with respect to the experiments directed at identifying the effect of different learning methods on classification performance: (i) Decision Tree (C4.5, J48 in WEKA), (ii) Naive Bayes Tree (NBTree), (iii) Naive Bayes (NaiveBayes), (iv) Averaged One-Dependence Estimators (AODE), (v) Bayesian Network (BayesNet), (vi) Radial Basis Function Network (RBF Network), (vii) Sequential Minimal Optimisation (SMO), (viii) Neural Network (Multilayerperceptron) and (ix) Logistic Regression (Logistic). In each case the Combined data, $K = 15$ and information gain feature selection was used. The results are presented in Table 5. From the table it can clearly be observed that Bayesian Network learning outperformed all the other classifier generator algorithms considered. The C4.5 decision tree generator produced substantially the worst performance.

5 Conclusion

In this paper a framework for remotely collecting census data using satellite imagery and data mining (classification) has been described. The main idea presented in this paper is that classifiers can be built that classify household satellite images to produce census data, provided an appropriate representation is used. The proposed representation is founded on the idea of representing segmented households using a histogram based formalism. The proposed framework was evaluated using differ-

Table 5 Classification performance using different learning methods (Combined, Information Gain and $K = 15$).

Learning methods	Accu-racy	Sensi-tivity	Speci-ficity	Precision "Small"	Precision "Large"	F-Measure	ROC Area
C4.5	0.600	0.765	0.385	0.619	0.556	0.684	0.643
Naive Bayes Tree	0.900	1.000	0.769	0.850	1.000	0.919	0.946
Naive Bayes	0.900	1.000	0.769	0.850	1.000	0.919	0.946
AODE	0.800	0.882	0.692	0.789	0.818	0.833	0.846
Bayesian Network	0.933	1.000	0.846	0.895	1.000	0.944	0.959
RBF Network	0.833	0.882	0.769	0.833	0.833	0.857	0.869
SMO	0.900	0.941	0.846	0.889	0.917	0.914	0.894
Neural Network	0.900	0.941	0.846	0.889	0.917	0.914	0.964
Logistic Regression	0.833	0.824	0.846	0.875	0.786	0.848	0.964

ent data sets, feature selection mechanisms, numbers of attributes (K) and learning methods. The reported results demonstrate that the best performance was produced using: (i) the Combined representation that combined histogram bin data with statistical information derived from the histograms, (ii) information gain as the feature selection method, (iii) $K = 15$ and (iv) Bayesian Network learning. The best results obtained were a sensitivity of 1.000 and a specificity of 0.846; these are extremely good results indicated that accurate census data can be collected using the proposed approach at a significantly reduced overall cost compared to traditional approaches to collecting such data.

References

1. Otto Bretscher. *Linear Algebra with Applications (3rd Edition)*. Prentice Hall, July 2004.
2. C. R. Brice and C. L. Fennema. Scene Analysis Using Regions. *Artificial Intelligence*, 1(3):205–226, 1970.
3. John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679 –698, nov. 1986.
4. Jin-Song DENG, Ke WANG, Jun LI, and Yan-Hua DENG. Urban land use change detection using multisensor satellite images. *Pedosphere*, 19(1):96 – 103, 2009.
5. Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, January 1972.
6. J. Faichney and R. Gonzalez. Combined colour and contour representation using anti-aliased histograms. In *Signal Processing, 2002 6th International Conference on*, volume 1, pages 735 – 739 vol.1, aug. 2002.
7. Juan A. X. Fanoe. Lessons from census taking in south africa: Budgeting and accounting experiences. *The African Statistical*, 13(3):82–109, 2011.
8. Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice Hall, 3 edition, August 2007.
9. E. L. Hall. Almost uniform distributions for computer image enhancement. *IEEE Trans. Comput.*, 23(2):207–208, February 1974.

10. Wynne Hsu, S. T. Chua, and H. H. Pung. An integrated color-spatial approach to content-based image retrieval. In *Proceedings of the third ACM international conference on Multimedia*, MULTIMEDIA '95, pages 305–313, New York, NY, USA, 1995. ACM.
11. Jacek Kozak, Christine Estreguil, and Katarzyna Ostapowicz. European forest cover mapping with high resolution satellite data: The carpathians case study. *International Journal of Applied Earth Observation and Geoinformation*, 10(1):44 – 55, 2008.
12. Tzu-Chuen Lu and Chin-Chen Chang. Color image retrieval technique based on color features and image bitmap. *Inf. Process. Manage.*, 43(2):461–472, March 2007.
13. G. Mallinis, I.D. Mitsopoulos, A.P. Dimitrakopoulos, I.Z. Gitas, and M. Karteris. Local-scale fuel-type mapping and fire behavior prediction by employing high-resolution satellite imagery. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 1(4):230 –239, dec. 2008.
14. V. Mesev. The use of census data in urban image classification. *PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING*, (5):431–438, May 1998.
15. F. O'Gorman and M. B. Clowes. Finding picture edges through collinearity of feature points. *IEEE Trans. Comput.*, 25(4):449–456, April 1976.
16. Nobuyuki Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.
17. N. Papamarkos and B. Gatos. A new approach for multilevel threshold selection. *CVGIP: Graph. Models Image Process.*, 56(5):357–370, September 1994.
18. T. Pavlidis. *Algorithms for Graphics and Image Processing*. Springer, 1982.
19. K. Roychowdhury, S. Jones, C. Arrowsmith, and K. Reinke. Indian census using satellite images: Can dmsp-ols data be used for small administrative regions? In *Urban Remote Sensing Event (JURSE), 2011 Joint*, pages 153 –156, april 2011.
20. Xiang-Yang Wang, Jun-Feng Wu, and Hong-Ying Yang. Robust image retrieval based on color histogram of local feature regions. *Multimedia Tools Appl.*, 49(2):323–345, August 2010.
21. Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 3. edition, 2011.
22. Yang Zhang, Rongyi He, and Muwei Jian. Comparison of two methods for texture image classification. In *Proceedings of the 2009 Second International Workshop on Computer Science and Engineering - Volume 01*, IWCSE '09, pages 65–68, Washington, DC, USA, 2009. IEEE Computer Society.
23. S. W. Zucker and D. Terzopoulos. Finding structure in co-occurrence matrices for texture analysis. *Computer Graphics and Image Processing*, 12:286–308, 1980.
24. Reyer Zwiggelaar. Texture based segmentation: Automatic selection of co-occurrence matrices. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1 - Volume 01*, ICPR '04, pages 588–591, Washington, DC, USA, 2004. IEEE Computer Society.