Satellite Image Mining for Census Collection: A Comparative Study With Respect to the Ethiopian Hinterland

Kwankamon Dittakan[1], Frans Coenen[1], and Rob Christley[2]

[1] Department of Computer Science,
University of Liverpool, Liverpool, L69 3BX, United Kingdom
[2] Department of Epidemiology and Population Health,
University of Liverpool, Leahurst Campus,
Chester High Road, CH64 7TE Neston, Cheshire,United Kingdom
{dittakan,coenen,robc}@liverpool.ac.uk

Abstract. Census data provides an important source of information with respect to decision makers operating in many different fields. However, census collection is a time consuming and resource intensive task. This is especially the case in rural areas where the communication and transportation infrastructure is not as robust as in urban areas. In this paper the authors propose the use of satellite imagery for census collection. The proposed method is not as accurate as "on ground" census collection, but requires very little resource. The proposed method is founded on the idea of collecting census data using classification techniques applied to relevant satellite imagery. The objective is to build a classifier that can label households according to "family" size. More specifically the idea is to segment satellite images so as to obtain pixel collections describing individual households and represent these collections using some appropriate representation to which a classifier generator can be applied. Two representations are considered, histograms and Local Binary Patterns (LBPs). The paper describes the overall method and compares the operation of the two representation techniques using labelled data obtained from two villages lying some 300km to the northwest of Addis Ababa in Ethiopia.

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

1 Introduction

National census data provides an important source of statistical information with respect to planners and decision makers working in a wide diversity of domains. Census data is seen as an important source of information with which to measure the "well being" of a population and to provide input to national and regional development projects (both economic and social). For example development of public services such as education, health and transport services. Census data is typically obtained using a questionnaire format either for self-completion by individuals (in which case they are typically distributed by post or electronically), or for completion by field staff. There are a number

of obstacles to the collection of national census data: it is both time consuming and resource intensive (with respect to both collection and analysis) while at the same time people are often reluctant to participate (typically they are suspicious of the motivation for the census). In financial terms the cost of census collection is often substantial. For example it has been suggested that the 2006 Australian national census cost an estimated $300 million Australian dollars, while the 2010 US census was expected to cost more than $11 billion US dollars and involve a million part time employees[3]. The difficulties associated with census collection are compounded in rural areas where the population tends to be sparser and the communication and transport infrastructure tends to be less robust than in rural areas [8].

The solution proposed in this paper is founded on the idea of using satellite imagery to generate census data by segmenting images, identifying households and using a classifier to predict household size (number of people living in the household). Given a training set of hand-labeled households we can build a classifier to predict household type and size and use this to generate census information. The proposed approach is not applicable with respect to all areas (such as inner city areas where population estimates are difficult to obtain from satellite imagery) but is applicable in more rural areas. The focus for the study is the Ethiopia hinterland. The advantages offered by the proposed approach are: (i) low cost, (ii) speed of collection and (iii) automate processing. The disadvantage is that it will not be as accurate as more traditional "on ground" census collection, however it is suggested that the advantages out weigh the disadvantages.

The main challenge of the proposed census collection method, with respect to the work presented in this paper, is how best to represent the image data so that classifier generation techniques can be applied and census data effectively collected. Two image representations are considered: (i) Colour Histograms and (ii) Local Binary Patterns (LBPs). These are two techniques that have been "tried and tested" with respect to other image analysis applications. Histogram representations have been widely used for *whole image* representation (see for example [11]) because they offer the advantage that they obviate the need for image object identification. LBPs, alternatively, have been extensively used with respect to texture analysis [16].

The proposed approach is fully described in the remainder of this paper and evaluated using test data collected from two villages lying some 300km to the northwest of Addis Ababa in Ethiopia. The rest of this paper is organised as follows. In Section 2 some previous works is presented. Section 3 provides detail of the geographical study area in Ethiopia used for evaluation purposes. Section 4 then provides a description of the proposed census mining framework. Section 5 describes the proposed colour histogram based representation and Section 6 the proposed LBP representation. Section 7 reports on the evaluation of the framework. Finally, a summary and some conclusions are presented in Section 8.

2 Previous Work

Image understanding is an important and fundamental problem in domains such as computer vision and pattern recognition where the main objective is to understand the char-

[3] http://usgovinfo.about.com/od/censusandstatistics/a/aboutcensus.htm

acteristics of an image and interpret its semantic meaning. Image classification is an emerging image interpretation technique which can be used to categorise image sets according to a predefined set of labels. The performance of classifiers depends on the quality of the features used, features such as colour and texture. One method of encapsulating image colour is to use a histogram image representation technique whereby colour histograms represents the number of pixels associated with a particular subset of colours. The advantages offered by histogram-based representation are: (i) low storage requirements, (ii) automated generation and (iii) fast querying. Histogram representations have been used with respect to many applications including image retrieval [19, 4, 13] and remote sensing applications such as land usage analysis [1, 2] and land change detection [14, 3]. The histogram representation is one of the representations considered in this paper.

Texture is an important feature with respect to both human and computer vision. one example where texture analysis has been usefully employed is with respect to pattern recognition [21]. There are three principle mechanisms that may be adopted to describe the texture in digital images: (i) statistical, (ii) structural and (iii) spectral. The statistical approach is concerned with capturing texture using quantitative measures such as "smooth", "coarse" and "grainy". Structural approaches describe image texture in terms of a set of texture primitives or elements (texels) that occur as regularly spaced or repeating patterns. In the spectral approach the image texture features are extracted by using the properties of (say) the Fourier spectrum domain so that "high-energy narrow peaks" in the spectrum can be identified [9]. Local Binary Patterns (LBPs) are a texture representation method which is both statistical and structural in nature [17]. Using the LBP approach a binary number is produced, for each pixel, by thresholding its value with its neighbouring pixels. LBP offers the advantages of tolerance with respect to illumination changes and its computational simplicity. The LBP method has been used in many application such as face recognition [10, 20]. The LBP representation is the second representation considered in this paper.

Remote Sensing is concerned with techniques for observing the Earths surface, or its atmosphere, using sensors located on spacecraft or aircraft platforms and producing images of regions on the earths surface as a result. Satellite image interpretation offers advantages with respect to many applications for example: geoscience studies, astronomy, military intelligence, and geographic information systems [6, 12]. There are a small number of reports available on the use of satellite imagery for census data collection purposes. For example "nightlight" satellite images have been used to produce population census data and to analysis issues concerned with population density at the "sub-district level" [5]. In [15] classification techniques were applied to satellite image data to estimate the population density distribution with respect to one kilometre "blocks". The difference between the work described in [15] and that proposed in this paper is that the considered approach operates at a much finer level of granularity. The authors have themselves conducted some previous work concerned with the application of classification techniques to satellite imagery to generate census data. This is described in [7]. The work attempted to define satellite image data using an earlier version of the histogram based approach presented in this paper, the evaluation was also directed at a much smaller data set and therefore not conclusive.

3 Case Study Application Domain

To act as a focus for the research a case study was considered directed at a rural area within the Ethiopian hinterland, more specifically two data sets were collected with respect to two villages (Site A and Site B) located within the Harro district in the Oramia Region of Ethiopia (approximately 300 km north-west of Addis Abba) as shown in Figure 1. Site A was bounded by the parallels of latitude 9.312650N and 9.36313N, and the meridians of longitude 37.123850E and 37.63914E and. Site B was bounded by the parallels of latitude 9.405530N and 9.450000N, and the meridians of longitude 36.590480E and 37.113550E. Using the know bounding latitudes and longitudes of our two test sites appropriate satellite imagery was extracted from Google Earth[4]. The images were originally obtained using the GeoEye satellite with a 50 centimetre ground resolution. The satellite images for Site A were released by Google Earth on 22 August 2009 (Figure 1(b)) and those for Site B (Figure 1(c)) on 11 February 2012. The Site B satellite images were obtained during the "dry season" (September to February), while the site A images were obtained during the rainy season (June to August). From Figure 1(b) the households can be clearly identified, many of the households have tin roofs which are easy to differentiate from the (green) backgrounds, the households are less easy to identify in Figure 1(a) where they tend to merge into the (light-brown) background. On-the-ground household data (including family size and household latitude and longitude) was collected by University of Liverpool field staff in May 2011 and July 2012. The minimum and maximum family size were 2 and 12 respectively, the mean was 6.31, the medium were 6 and standard deviation was 2.56. These two data sets then provided the training and test data required for our proposed census collection system.

4 Census Mining Framework

The proposed census mining framework is presented in this section. A schematic of the proposed framework is given in Figure 2. The framework supports a three phase census collection form satellite imagery process: (i) Data preprocessing, (ii) Classifier generation and (iii) Classifier evaluation.

During the data preprocessing phase (left hand block in Figure 2) the satellite image input data is prepared ready for the application of the classifier generation phase. The preprocessing stage comprises five individual stages: (i) coarse segmentation. (ii) image enhancement, (iii) detailed segmentation, (iv) image representation and (v) feature selection. During coarse segmentation the input imagery is roughly segmented to give a set of large scale sub-images each covering a number of households (typically between two and four). In the next stage various image enhancement processes are applied to the identified sub-images. During the detailed segmentation stage the enhanced coarse sub-images are segmented to obtain individual households. Figures 3(a) and (b) show two example of segmented household images taken from Site A and Site B respectively. The result is one image per household. For classifier generation purposes each labeled

[4] http://www.google.co.uk/intl/en_uk/earth/index.html

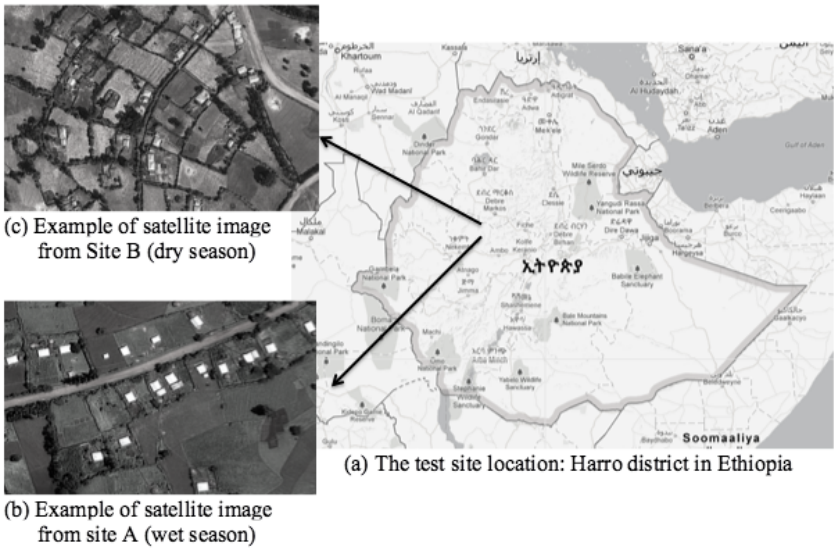

Fig. 1. The test site location: Harro district in Ethiopia

segmented households must be represented in a manner that ensures that the salient features are maintained (so as to ensure an effective classifier); as noted in the introduction to this paper two representation techniques are considered: a histogram based technique and an LBP based technique. The final step in the preprocessing phase comprised feature selection, the aim here was to reduce the overall size of the feature space (histogram based or LBP based) so that those features that best served to discriminate between classes were retained. For details concerning steps 1 to 3 the reader is referred to the authors earlier work presented in [7]. The histogram and LBP representations to support effective classifier generation are amongst the main contributions of this paper and are considered in more detail in Sections 5 and 6 respectively.

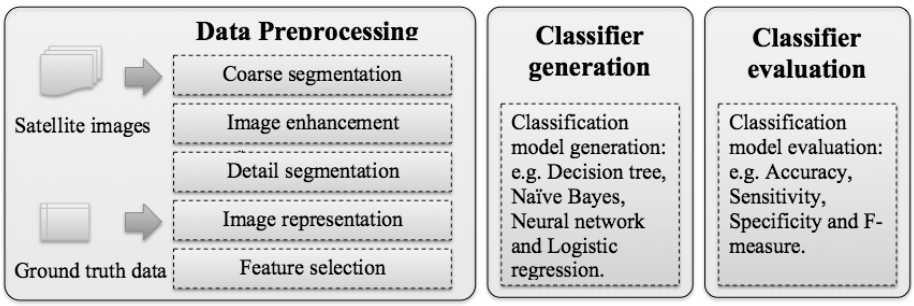

Fig. 2. Proposed census data from satellite imagery framework.

Classifier generation is the second phase (centre block in Figure 2) in the proposed framework during which the desired classifier was generated from labeled training data produced during the data preprocessing phase (Phase1) described above. The final phase (right hand block in Figure 2) was classifier evaluation where the classifier was applied to a labelled test set and the generated results compared with known results, the aim was to produce statistical measures indicating the confidence that can be associated with the generated classifier.

(a) An example household image from Site A (wet season)

(b) An example household image from Site B (dry season)

Fig. 3. Example segmented household images

5 Color Histogram

In the histogram based satellite image representation image colour is the central feature used. Image colour distribution is captured using a set of histograms (one set per satellite image). The advantage offered is that histograms are simple to generate, invariant to translation and rotation of image content, low on storage requirement and allow for fast query execution. The X-axis of each histogram comprises a number of "bins" each representing a "colour range". The Y axis of each histogram then represents the number of pixels falling into each bin. For each preprocessed household satellite image seven different histograms was extracted: (i) three histogram from the RGB colour channels (red, green, blue), (ii) three histograms from the HSV colour channels (hue, saturation, value) and (iii) a grayscale histogram. Each of the seven histograms comprised 32 bins, giving 224 (7×32) features in total. Figure 4 shows the seven example histograms produced from one of the identified household image used in the evaluation presented below (Section 7).

A simple alternative representation was to extract some simple statistical colour information from the image data. The idea here was that this statistical information could be used to augment the colour histogram information (or used as a representation

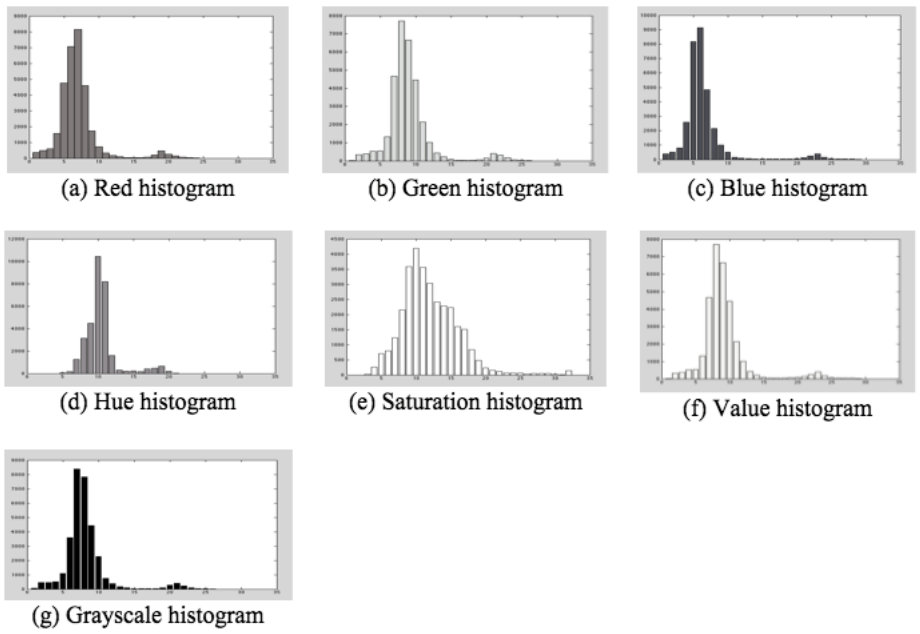

Fig. 4. Histogram representation for an example image.

on its own). A total of 13 statistical features were identified: (i) 5 features describing the RGB colour channels, (ii) 5 features describing the HSV colour channels and (iii) 3 feature describing the grayscale channel. The individual features are listed in Table 1. Thus on completion of the histogram based representation stage each household is represented using a feature vector of length 237 ($224 + 13 = 237$).

Table 1. Additional colour based statistical features.

#	RGB color channel descripton	#	HSV color channel description	#	Grayscale channel description
1	Average red	6	Average hue	11	Mean of grayscale
2	Average green	7	Average saturation	12	Standard deviation of grayscale
3	Average blue	8	Average value		
4	Mean of RGB	9	Mean of HSV	13	Average of grayscale histogram
5	Standard deviation of RGB	10	Standard deviation of HSV		

6 Local Binary Pattern

As already noted, Local Binary Patterns (LBPs) are generally used for representing image texture. However, there is no reason why LBPs cannot be used to represent images irrespective of whether we are interested in texture or not. The LBP representation offers the advantages that they are easy to generate and tolerant against illumination changes. The use of LBPs was therefore considered as an alternative to the proposed histogram based representation.

To generate a set of LBPs from individual household images the images were first transform into grayscale. A 3×3 pixel window, with the pixel of interest at the centre, was then used as the basic "neighbourhood" definition with respect to the LBP representation. For each neighbourhood the grayscale value for the centre pixel was defined as the threshold value with which the surrounding eight neighbourhoods were compared. For each neighbourhood pixel a 1 was recorded if the grayscale value of the neighbourhood pixel was greater than the threshold, and a 0 otherwise. The result is an eight digit binary number. In other words 256 (2^8) different patterns can be described (note that LBPs calculated in this manner are not rotation invariant).

Variations for the basic LBP concept can be produced by using different sizes (radii) of neighbourhoods. These variations can be described using the (P,R) notation where P is the number of sampling points and R is the radius surrounding the centre pixel [18]. For evaluation purposes three different variations of the LBP representation were used (Figure 5): LBP(8,1), 8 sampling points within a radius of 1; LBP(8,2), 8 sampling points within a radius of 2; and LBP(8,3), 8 sampling points within a radius of 3.

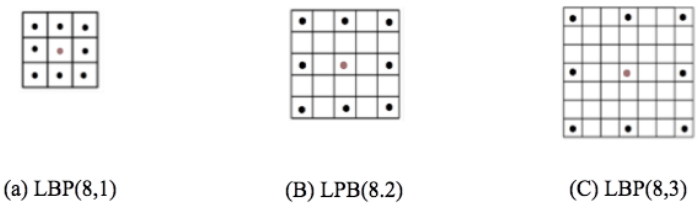

Fig. 5. Local Binary Pattern (LBP) variations.

The resulting LBP representation were conceptualised in terms of a 2^R dimensional feature vector where each element represented a potential LBP value and the value held within each element corresponded to the number of pixels associated with each LBP value.

As in the case of the histogram representation, an alternative to the LBP representation is to use statistical measures of texture. Again the idea was that such statistical features could be used to augment the LBP representation (or used as a representation on its own). Three categories of texture statistic were identified: (i) entropy features (E), (ii) grey-level occurrence matrix features (M) and (iii) wavelet transform features (W). Table 2 lists the statistical features generated (the letter in parenthesis in each

case indicates the category of the feature). Thus on completion of the LBP based representation stage each household is represented using a feature vector of length 266 ($2^R + 10 = 2^8 + 10 = 256 + 10 = 266$).

Table 2. Additional texture based statistical features.

#	Description	#	Description	#	Description
1	Entropy (E)	7	Average approximation coefficient matrix, *cA* (W)	10	Average diagonal coefficient matrix, *cD* (W)
2	Average Local Entropy (E)				
3	Contrast (M)	8	Average horizontal coefficient matrix, *cH* (W)		
4	Correlation (M)				
5	Energy (M)	9	Average vertical coefficient matrix, *cV* (W)		
6	Homogeneity (M))				

7 Evaluation

To evaluate and compare the proposed histogram and LBP representations in the context of classifier generation for census data collection the test data introduced in Section 3 was used. A total of 120 records were selected from the two test sites: 70 records from Site A and 50 records from Site B. The labeled household data was separated into three classes: (i) "small family" (less than or equal to 5 people), (ii) "medium family" (between 6 and 8 people inclusive) and (iii) "large family" (more than 8 people). Some statistics concerning the class distributions for the Site A and B data sets are presented in 3.

Table 3. Class label distribution for data set Site A and data set Site B

	small family	medium family	large family	total
Site A	28	32	10	70
Site B	19	21	10	50
total	47	53	20	120

Before classification could commence the input data was first discretised (ranged), a sub-process that served to decrease the size of the feature space (fewer values for each feature/dimension). Feature selection was then applied so as to reduce the number of dimensions (Step 5 in Phase 1 of the proposed framework). In the context of the evaluation presented in this section the Chi-square feature selection strategy was applied. For

classifier generation purposes a number of classifier learners were used as implemented in the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench[5]. For the evaluation purposes Ten fold Cross-Validation (TCV) was applied and the effectiveness of the generated classifiers reported in terms of: (i) accuracy (AC), (ii) area under the ROC curve (AUC), (iii) sensitivity (SN), (iv) specificity (SP) and precision (PR).

Three sets of experiments were conducted directed at:

1. A comparison of the proposed histogram and LPB based representations in the context of classification for census data collection.
2. Identification of the most appropriate number (K) of attributes (features) to retain during feature selection.
3. Determination of the most appropriate classifier generator to use (for the purpose of classification for census data collection).

each set of experiments is discussed in further detail in the following three subsections (Subsections 7.1, 7.2 and 7.3).

7.1 Colour histograms v. LBPs

To compare the operation of the proposed representations three different variations of the histogram representation were considered together with seven variations of the LBP representation. The histogram representations considered were: (i) colour histogram only (CH), (ii) colour statistics only (CS) and (iii) colour histogram and statistics combined (CH+CS). The LBP representations considered were: (i) LBP(8,1) only, (ii) LBP(8,2) only, (iii) LBP(8.3) only, (iv) texture statistics only (TS), (v) LBP(8,1) and statistics combined (LBP(8,1)+TS), (vi) LBP(8,2) and statistics combined (LBP(8,2)+TS) and (vii) LBP(83) and statistics combined (LBP(8,3)+TS). For the experiments Chi-Square feature selection was used with $K = 35$ (K = 25 was used for the experiments reported in Sub-section 7.2, had revealed that this was the most appropriate value for K) and a Neural Network learning method as these had been found to work well (see Sub-section 7.3). The results are presented in Table 4 (highest values indicated in bold font). Although the results are not conclusive from the Table it can be observed that:

- With respect to the colour histogram based representation best results were obtained using CH for Site A (wet season) and CH+CS for Site B (dry season), the distinction is assumed to occur because of the predominantly green colour of the wet season images in comparison with the predominantly brown colour of the dry season images.
- With respect to the LBP representation best results were produced using LBP(8,1) with respect to both Site A and Site B.
- Comparing the results obtained using both the colour histogram and the LBP representations, LBP(8,1) produced the best overall results.

Thus we conclude that in the context of the test scenario the LBP representation produced the most eftective results.

[5] http://www.cs.waikato.ac.nz/ml/weka/

Table 4. Comparison of different variations of the proposed histogram and LPB based representations in terms of classification performance (neural network classifier and $K = 35$)

Types	Data set	Site A					Site B				
		AC	AUC	PR	SN	SP	AC	AUC	PR	SN	SP
Histogram	CH	**0.614**	**0.754**	**0.615**	**0.614**	**0.761**	0.560	0.753	0.568	0.560	0.723
	CS	0.400	0.550	0.404	0.400	0.629	0.480	0.645	0.480	0.307	0.754
	CH+CS	0.557	0.741	0.551	0.557	0.713	**0.580**	**0.755**	**0.602**	**0.580**	**0.728**
LBP	LBP(8,1)	**0.771**	**0.880**	**0.758**	**0.771**	**0.856**	**0.600**	**0.792**	**0.599**	**0.600**	**0.764**
	LBP(8,2)	0.614	0.765	0.618	0.614	0.725	0.600	0.705	0.600	0.600	0.762
	LBP(8,3)	0.586	0.718	0.583	0.586	0.710	0.600	0.792	0.599	0.600	0.764
	TS	0.414	0.502	0.379	0.414	0.595	0.500	0.650	0.602	0.500	0.754
	LBP(8,1)+TS	0.757	0.848	0.754	0.757	0.847	0.600	0.768	0.601	0.600	0.757
	LBP(8,2)+TS	0.643	0.770	0.645	0.643	0.746	0.580	0.736	0.580	0.580	0.742
	LBP(8,3)+TS	0.557	0.706	0.553	0.557	0.686	0.600	0.768	0.601	0.600	0.757

7.2 Number of Attributes

In order to investigated the effect on classification performance of using different values of K with respect to Chi-Squared feature selection a sequence of experiments were conducted using a range of values for K from 10 to 35 incrementing in steps of 5. For the experiments the colour histogram (CH) and LBP(8,1) representations were used because previous experiments had indicated that these two representations produced the best performance (see above). Neural Network machine learning was again adopted. The results produced are presented in Table 5 (best values obtained are again highlighted in bold). From the Table the following can be observed:

- With respect to the colour histogram based representation best results tended to be obtained using $K = 25$ for Site A (wet season) and $K = 10$ for Site B (dry season), with better results being produced using the Site B data.
- With respect to the LBP representation best results tended to be produced using $K = 35$ with respect to both the Site A and the Site B data.
- The LBP(8,1) representation outperformed the Colour histogram (CH) representation.

Thus it can be concluded that higher K values, such as $K = 35$, produce better results using the LBP representation, while with respect to the histogram representation lower values for K ($K = 10$) produced the best result. Here it should be noted that as K increases the time complexity increases. For example the processing time for LBP(8,1) with respect to the Site A data set using a neural network learning method with $K = 15$, $K = 25$ and $K = 35$ was 6.07, 16.47 and 32.73 seconds respectively.

7.3 Learning Methods

Eight learning method were considered with respect to the experiments directed at identifying the effect of different learning methods on classification performance including:

Table 5. Comparison of different values of K with respect to Chi-Squared feature selection in terms of classification performance (CH and LBP(8,1), and neural network classifier)

Data set	No of attribute	Site A					Site B				
		AC	AUC	PR	SN	SP	AC	AUC	PR	SN	SP
CH	10	0.574	0.678	0.576	0.571	0.691	**0.680**	**0.777**	**0.679**	**0.680**	**0.801**
	15	0.614	0.719	0.608	0.614	0.754	0.580	0.757	0.559	0.580	0.749
	20	0.586	0.727	0.588	0.586	0.734	0.640	0.780	0.645	0.640	0.776
	25	**0.629**	**0.766**	**0.631**	**0.629**	**0.768**	0.620	0.796	0.637	0.620	0.754
	30	0.614	0.776	0.607	0.614	0.751	0.600	0.773	0.608	0.600	0.752
	35	0.614	0.754	0.615	0.614	0.761	0.560	0.753	0.568	0.560	0.723
LBP(8,1)	10	0.657	0.786	0.646	0.657	0.772	0.580	0.677	0.587	0.580	0.745
	15	0.757	0.875	0.752	0.757	0.835	0.580	0.699	0.590	0.580	0.730
	20	0.714	0.874	0.719	0.714	0.835	0.580	0.718	0.593	0.580	0.725
	25	0.757	0.887	0.749	0.757	0.849	0.600	0.713	0.595	0.600	0.754
	30	0.757	0.868	0.743	0.757	0.837	0.580	0.761	0.581	0.580	0.742
	35	**0.771**	**0.880**	**0.758**	**0.771**	**0.856**	**0.600**	**0.792**	**0.599**	**0.600**	**0.764**

(i) Decision Tree generators (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Networks (RBF Networks), (vi) Logistic Regression, (vii) Sequential Minimal Optimisation (SMO) and (viii) Neural Networks Back-propagation (in WEKA this is referred to as a MultilayerPerceptron). The Colour Histogram (CH) and LBP(8,1) representations were again used. $K = 10$ Chi-Squared feature selection was used with the Colour Histogram (CH) representation and $K = 35$ for the LBP(8,1) representation. The obtained results are presented in Table 6. From the Table it can be observed that:

- With respect to the colour histogram based representation best results were obtained using the Bayes Network learner with respect to the Site A data, and AODE and Bayes Network with respect to the Site B data.
- With respect to the LBP based representation best results were obtained using using Neural Networks with respect to the Site A data, and Logistic Regression with respect to the Site B data.
- The C4.5, Naive Bayes, RBF Network and SMO learners did not perform well.
- Overall the Neural Network learner, combined with the LBP(8,1) representation and $K = 35$ Chi-Squared feature selection, produced the best overall result.

Thus in conclusion a number of different machine learners produced good results, different machine learners tended to be more compatible with different representations, but overall Neural Network learning produced the best result.

8 Conclusion

A data mining mechanism for generating census data (household size) from satellite imagery has been proposed. More specifically a three phase framework has been suggested

Table 6. Comparison of different classifier generators in terms of classification performance (CH with $K = 10$ and LBP(8,1) with $K = 35$)

Data set	Learning method	Site A					Site B				
		AC	AUC	PR	SN	SP	AC	AUC	PR	SN	SP
CH + 35 atts	C4.5	0.557	0.640	0.573	0.557	0.677	0.480	0.618	0.450	0.480	0.681
	Naive Bayes	0.629	0.731	0.635	0.629	0.734	0.640	0.780	0.646	0.640	0.797
	AODE	0.586	0.709	0.605	0.586	0.686	**0.700**	0.706	**0.702**	**0.700**	0.813
	Bayes Network	**0.629**	**0.751**	**0.630**	**0.629**	**0.736**	0.680	**0.794**	0.684	0.680	**0.815**
	RBF Network	0.571	0.692	0.576	0.571	0.720	0.440	0.649	0.449	0.440	0.700
	Logistic Regression	0.486	0.637	0.486	0.486	0.670	0.580	0.700	0.560	0.580	0.744
	SMO	0.457	0.542	0.481	0.457	0.598	0.600	0.710	0.589	0.600	0.768
	Neural Network	0.574	0.678	0.576	0.5710	.691	0.680	0.777	0.679	0.680	0.801
LBP(8,1) + 35 atts	C4.5	0.614	0.718	0.619	0.614	0.773	0.500	0.657	0.495	0.500	0.708
	Naive Bayes	0.543	0.709	0.583	0.543	0.763	0.540	0.762	0.594	0.540	0.796
	AODE	0.557	0.755	0.529	0.557	0.698	0.600	0.782	0.596	0.600	0.766
	Bayes Network	0.571	0.729	0.599	0.571	0.768	0.520	0.767	0.554	0.520	0.772
	RBF Network	0.657	0.740	0.665	0.657	0.818	0.600	0.743	0.601	0.600	0.783
	Logistic Regression	0.757	0.857	0.757	0.757	0.866	**0.757**	**0.857**	**0.758**	**0.757**	**0.866**
	SMO	0.729	0.789	0.718	0.729	0.823	0.729	0.789	0.718	0.729	0.823
	Neural Network	**0.771**	**0.880**	**0.758**	**0.771**	**0.856**	0.600	0.792	0.599	0.600	0.764

that takes large scale satellite imagery as input and produces an evaluated classifier for household size census collection. The key element with respect to the process is the way in which individual households are represented so that an effective classifier can be generated. Two basic representations were proposed: colour histograms and LBPs. Ten variations of these representations were considered. Experiments were conducted using a training set obtained from a rural part of the Ethiopian hinterland. This was selected because of the availability of on-the-ground data and because the proposed process is intended for use in rural areas, particularly rural areas with poor communication and transport infrastructures that tend to exacerbated the issues associated with traditional forms of census collection. Experiments were also conducted to identify a best K value with respect to Chi-Squared feature selection and a number of classifier generators. The main findings were: (i) that it is possible to collect household size census data using the proposed approach to a reasonable level of accuracy (a best ROC value of 0.880 was obtained), (ii) that there was a performance distinction between the "green" wet season data (Site A) and the "brown" dry season data (Site B), (iii) that the LBP(8,1) representation tended to produce the best results, (iv) the most desirable value for K depended on the nature of the representation adopted (high with respect to the LBP representation, lower with respect to the histogram representation), and (v) that a number of machine learners performed well but the use of neural networks provided the best results. For future work the research team intend to investigate representations based on quad tree decompositions and better segmentation techniques to generate individual house hold images.

References

1. R. Abdelfattah and J.M. Nicolas. Interferometric synthetic aperture radar coherence histogram analysis for land cover classification. In *Proc. Information and Communication Technologies (ICTTA'06)*, volume 1, pages 343–348. IEEE, 2006.
2. P.M. Atkinson. Super-resolution land cover classification using the two-point histogram. In Xavier Sanchez-Vila, Jesus Carrera, and JosJaime Gmez-Hernndez, editors, *Proc. Geostatistics for Environmental Applications (geoENV'VI)*, volume 13, pages 15–28. Springer, Netherlands, 2004.
3. C. Beumier and M. Idrissa. Building change detection by histogram classification. In *Proc. Signal-Image Technology and Internet-Based Systems (SITIS'11)*, pages 409–415. IEEE, 2011.
4. R. Chakravarti and Xiannong Meng. A study of color histogram based image retrieval. In *Proc. Information Technology: New Generations (ITNG'09)*, pages 1323–1328. IEEE, 2009.
5. L. Cheng, Y. Zhou, L. Wang, S. Wang, and C. Du. An estimate of the city population in china using dmsp night-time satellite imagery. In *Proc. IEEE International on Geoscience and Remote Sensing Symposium (IGARSS'07)*, pages 691–694. IEEE, 2007.
6. H. Demirel and G. Anbarjafari. Satellite image resolution enhancement using complex wavelet transform. *Geoscience and Remote Sensing Letters, IEEE*, 7(1):123–126, 2010.
7. K. Dittakan, F. Coenen, and R. Christley. Towards the collection of census data from satellite imagery using data mining: A study with respect to the ethiopian hinterland. In Max Bramer and Miltos Petridis, editors, *Proc. Research and Development in Intelligent Systems XXIX*, pages 405–418. Springer, London, 2012.
8. J. A. X. Fanoe. Lessons from census taking in south africa: Budgeting and accounting experiences. *The African Statistical*, 13(3):82–109, 2011.
9. R. C. Gonzalez and R. E. Woods. *Digital Image Processing (3rd Edition)*. Pearson Prentice Hall, 3 edition, 2007.
10. A. Hadid. The local binary pattern approach and its applications to face analysis. In *Proc. first workshop on Image Processing Theory, Tools and Applications (IPTA'08)*, pages 1–9. IEEE, 2008.
11. M.H.A. Hijazi, F. Coenen, and Y. Zheng. Retinal image classification using a histogram based approach. In *Proc. International Joint Conference on Neural Network (Special Session on Soft Computing in Medical Imaging), part of IEEE World Congress on Computational Intelligence (WCCI'10)*, pages 3501–3507, 2010.
12. J. H. Jang, S. D. Kim, and J. B. Ra. Enhancement of optical remote sensing images by subband-decomposed multiscale retinex with hybrid intensity transfer function. *Geoscience and Remote Sensing Letters, IEEE*, 8(5):983–987, 2011.
13. S. Jeong, C. S. Won, and R. M. Gray. Image retrieval using color histograms generated by gauss mixture vector quantization. *Computer Vision and Image Understanding*, 94(1-3):44–66, 2004.
14. Y. Kita. A study of change detection from satellite images using joint intensity histogram. In *Proc. International Conference on Pattern Recognition (ICPR'08)*, pages 1–4. IEEE, 2008.
15. G. Li and Q. Weng. Using landsat etm+ imagery to measure population density in indianapolis, indiana, usa. *Photogrammetric engineering and remote sensing*, 71(8):947, 2005.
16. T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
17. M. Pietikäinen. Image analysis with local binary patterns. In *Proc. Scandinavian conference on Image Analysis (SCIA'05)*, pages 115–118. Springer-Verlag Berlin, Heidelberg, 2005.

18. C. Song, P. Li, and F. Yang. Multivariate texture measured by local binary pattern for multispectral image classification. In *Proc. IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS'06)*, pages 2145–2148, 2006.
19. N. Vasconcelos and A. Lippman. Feature representations for image retrieval: beyond the color histogram. In *Proc. IEEE International Conference on Multimedia and Expo (ICME'00)*, volume 2, pages 899–902. IEEE, 2000.
20. G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'07)*, 29(6):915–928, 2007.
21. G. Zhao, G. Wu, Y. Liu, and J. Chen. Texture classification based on completed modeling of local binary pattern. In *Proc. International Conference on Computational and Information Sciences (ICCIS'11)*, pages 268–271. IEEE, 2011.