Population Estimation Mining Using Satellite Imagery

Kwankamon Dittakan, Frans Coenen, Rob Christley, and Maya Wardeh

Department of Computer Science,
University of Liverpool, Liverpool, L69 3BX, United Kingdom
`{dittakan,coenen,robc,maya.wardeh}@liverpool.ac.uk`

Abstract. Many countries around the world regularly collect census data. This census data provides statistical information regarding populations to in turn support decision making processes. However, traditional approaches to the collation of censes data are both expensive and time consuming. The analysis of high resolution satellite imagery provides a useful alternative to collecting census data which is significantly cheaper than traditional methods, although less accurate. This paper describes a technique for mining satellite imagery, to extract census information, founded on the use of classification techniques coupled with a graph based representation of the relevant imagery. The fundamental idea is to build a classifier that can label households according to "family size" which can then be used to collect census data. To act as a focus for the work training data obtained from villages lying some 300km to the northwest of Addis Ababa in Ethiopia was used. The nature of each household in the segmented training data was captured using a tree-based representation. Each tree represented household had a "family size" class label associated with it. This data was then used to build a classifier that can be used to predict household sizes according to the nature of the tree-based structure.

Keywords: Satellite Image Analysis and Mining, Data Mining Applications, Population Estimation Mining

1   Introduction

The work described in this paper is directed at the automated estimation of census information using data mining techniques applied to satellite imagery. The motivation for the work is that the collection of census data, when conducted in the traditional manner (using postal, email or interview approaches) is very resource intensive. This is especially the case in rural areas that lack sophisticated communication and transport infrastructure. The solution proposed in this paper is found on the concept of using satellite imagery for population estimation. The idea is to use a small sample of satellite images of households, where the "family size" is known, to build a classifier that can then be used to predict household family sizes over a much wider area. The main issue to be addressed is how best to represent the household image data so that classification techniques can be applied. The solution presented in this paper is to first segment relevant satellite imagery so as to isolate individual households and represent individual household data using a quad-tree based technique. The advantages offered by the proposed approach, in the context of census collection, are: (i) low cost, (ii) speed of

collection and (iii) automated 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 outweigh the disadvantages.

The proposed approach is more applicable with respect to rural areas than suburban and inner city areas. In this paper, the study area used as an exemplar application area is in the Ethiopia hinterland. More specifically training data obtained from two villages lying some 300 km to the northwest of Addis Ababa in Ethiopia was used, as shown in Figure 1 (the letters 'A' and 'B' indicate the village locations)[1].

The rest of this paper is organised as follows. In Section 2 some related work is briefly presented. Section 3 then provides a description of the proposed census mining framework. A brief overview of the proposed image segmentation process is presented in Section 4. In Section 5 details of the graph-based representation are presented, including a review of the proposed: quadtree decomposition, frequent subgraph mining and feature vector representation. The performance of the proposed census mining framework, using the Ethiopian test data, is then considered in Section 6. Finally, Section 7 provides a summary and some conclusions.

Fig. 1. Test site location.

2   Previous Work

From the literature there have been a number of reports concerning automatic census collection founded on a variety of technologies. For example, in [3] voice recognition was used to automatically collect census data using telephone links. Another example can be found in [10] where census data was automatically recorded using PDAs. These two reported methods both demonstrated that time savings can be gained by at least partially automating the census gathering process.

Many methods for population estimation have been reported in the literature. These can be categorised as being founded on either: (i) areal interpolation or (ii) statistical modelling. The areal interpolation approach is typically used to identify areas of

---

[1] http://maps.google.com

differing population densities, in other words to produce corse population density studies. The statistical modelling approach is typically used for identifying relationships between populations and other information sources such as Geographic Information System sources [12].

The approach advocated in tis paper is founded on the use of satellite imagery. Satellite imagery has been used with respect to population estimation. For example, Google Earth satellite images have been used to estimate corse population densities at the city and village levels [7]. By identifying features such as dwelling units and residential areas satellite images have also been applied for the purpose of population estimation [1]. Further examples can be found in [2, 8], where "night satellite" imagery was used to estimate population sizes according to the local densities of light sources.

In the context of the proposed population estimation mining the mechanism whereby satellite images are represented is important. There are many representation techniques available founded on image features such as: (i) colour, (ii) texture and (iii) structure. Colour histograms are wildly used to represent image content in terms of colour distribution. There are two major methods for histogram generation: the "binning" histogram and the clustering methods. The binning histogram method is used to generate a histogram by dividing the entire colour space into a number of bins and then, for each bin, recording the number of pixels that "fall into" that bin. Using the clustering method the colour space is first divided into a large number of bins and then a clustering algorithm is used to group them together [11]. The colour histogram representation has the advantages that it is easy to process and is invariant to translation and rotation of the image content [6]. Texture features can be used to describe a variety of surface characteristics of images [14]. There are three principle mechanisms that may be adopted to describe texture: statistical, structural and spectral. The statistical approach is concerned with capturing texture using quantitative measures such as "smooth", "coarse" and "grainy". The structural approaches describes 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 using properties of (say) the Fourier spectrum domain so that "high-energy narrow peaks" in the spectrum can be identified [5]. Structure features are used to describe the "geometry" of an image according to the relative position of elements that may be contained in an image. A well know structural image feature representation is the quad tree representation. This is then the fundamental representation used with respect to the work described in this paper.

3   Census Mining Framework

An overview of the proposed process for census mining is presented in this section. A schematic of the framework is shown in Figure 2. The framework comprised two phases (as represented by the rectangular boxes): (i) Preprocessing and (ii) Classification.

The first phase of the census mining framework is the preprocessing phase (the left rectangle in Figure 2) where the input data is prepared. The required preprocessing consists of two steps: image segmentation and image representation. The input is a satellite image of a given area covering (say) a number of villages. This image is then segmented in order to identify a set of individual households. This segmentation process
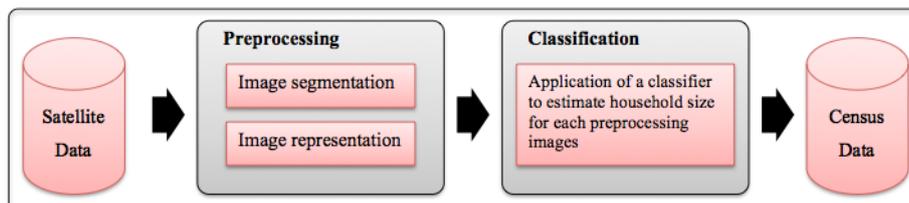
Fig. 2. Proposed census mining from satellite imagery framework

was described in detail in [4]; however, for completeness, a brief overview of the process is presented in Section 4. Next the identified household pixel data is translated into a representation that allows for the application of a classifier. In this paper a novel graph-based representation technique is proposed, detail of which is presented in Section 5.

After the households have been segmented and appropriately represented the classification phase may be commenced. This is relatively straight forward once an appropriate classifier has been generated. There are many classifier generation techniques that may be adopted and some of these are considered with respect to the evaluation presented in Section 6.

4   Segmentation

This section presents a brief overview of the image segmentation process as applied to the input data. The image segmentation comprises three individual stages: (i) coarse segmentation, (ii) image enhancement and (iii) fine segmentation. The first stage is thus coarse segmentation whereby the input satellite imagery is roughly separated into a set of sub-images covering (typically) between one and four households each. Once the coarse segmentation process is completed, the next stage is image enhancement where a range of image enhancement processes are applied to the coarse segmented sub-images so as to facilitate the following fine segmentation of individual households. During the fine segmentation stage, the enhanced coarse segmented sub-images are segment further so as to isolate individual households so that we end up with one image per household. Figure 3(a) and (b) show two fine segmented household images taken from test Site A and B respectively (see Figure 1).

The segmentation process is completed by translating the resulting RGB image data into a grayscale format ready for further processing followed by the application of a histogram equalisation process. Histogram equalisation is concerned with contrast adjustment using image histograms. Figure 4(a) sows the result when histogram equalisation is applied to the household image presented in Figure 3(a). For the purpose of the hierarchical quad tree decomposition (see below) a binary image transformation is then applied as shown in Figure 4(b).

5   Graph-Based Representation

Once a set of households has been fine segmented the next stage of the data preparation phase is to translate the segmented pixel data into a form suitable for the application
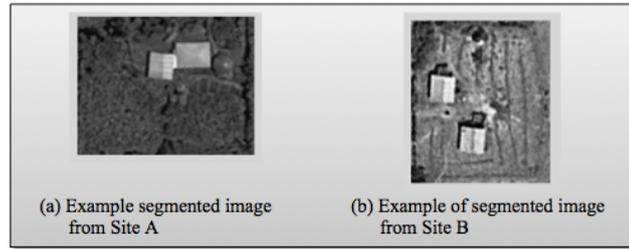
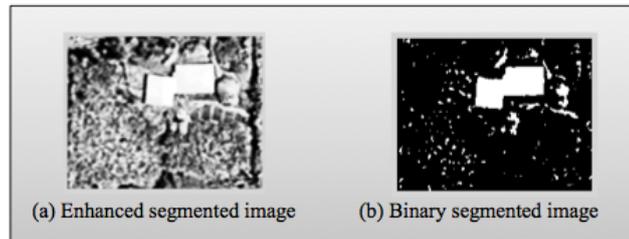**Fig. 3.** Example of segmented household from Site A and Site B

Fig. 4. Histogram equalisation and binary image transformation

of a classifier. The translation needs to be conducted in such a way that all salient information is retained while at the same time ensuring that the representation is concise enough to allow for effective further processing. The fundamental idea here is to adopt a graph based representation, more specifically a quad tree based representation (one per household). Quadtrees have been used extensively in the context of image processing (see for example [9]). However, the quadtree representation does not lend itself to ready incorporation with respect to classification algorithms. To do this we propose applying sub-graph mining to the the quadtree data to identify frequently occurring patterns across the data that can be used as features in the context of a feature vector representation. The patterns of interest are thus frequently occurring sub graphs. An overview of the process is presented in Figure 5. The graph-based representation consists four steps: (i) quadtree decomposition, (ii) tree construction, (iii) frequent subgraph mining and (iv) feature vector transformation.

The first step, the quadtree decomposition, commences by "cropping" each household image so that it is turned into a 128×128 pixel square image surrounding the main building comprising the household (this is automatically identifiable because it is the largest contiguous "white" region). The image was then recursively quartered into "tiles", as shown in Figure 6, until either: (i) uniform tiles (quadrants) were arrived at or (ii) a maximum level of decomposition was reached. Figure 6(a) shows an example preprocessed household image and the associated quadtree decomposition in Figure 6(b). The generated decomposition was then stored in a quad tree format. The nodes
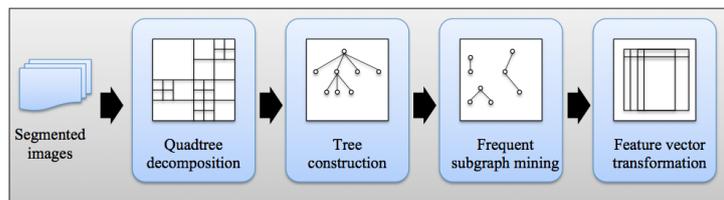
Fig. 5. Schematic illustration of the graph-based image representation processes.
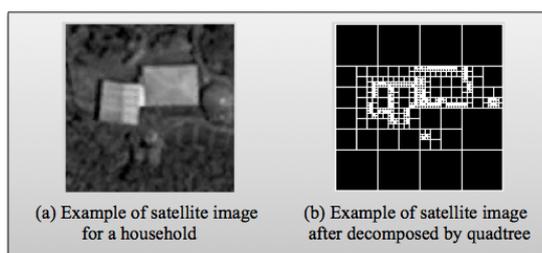

Fig. 6. The example of quadtree decomposition.

in this tree were labelled with a grayscale encoding generated using a mean histogram of grayscale colours for each block, in this manner eight labels were derived each describing a range of 32 consecutive intensity values. Figure 7 presents an example of a quad tree where the top level node (the root) represents the entire (cropped) image, the next level (Level 1) its immediate child nodes, and so on. In the figure the nodes are labelled numerically from 1 to 8 to indicate the grayscale ranges.The edges are labelled using a set of identifiers $\{1, 2, 3, 4\}$ representing the NW, NE, SW and SE child tiles associated with the decomposition of a particular parent tile. In Figure 7 the number in square brackets alongside each node is a unique node identifier derived according to the decomposition.

The quad tree (graph) based representation served to capture the content of individual fine segmented household images, although a disadvantage of the representation is the "boundary problem" where objects of interested may be located at the intersection of a decomposition. A second disadvantage is that the quad tree representation is not well suited to the purpose of classifier generation and subsequent usage of the generated classifier. The idea was therefore to identify frequently occurring patterns (subgraphs or subtrees) and treat these patterns as features within a feature vector representation. The motivation was the conjecture that such patterns would be indicative of commonly occurring features that might occur across the image set which in turn might be indicative of individual class labels. A number of different frequent subgraph miners could have been used; however, for the experiments described later in this paper, the well known gSpan frequent subgraph mining algorithm [13] was adopted. This uses the concept of a support threshold σ to define the concept of a frequent subgraph, the lower the value
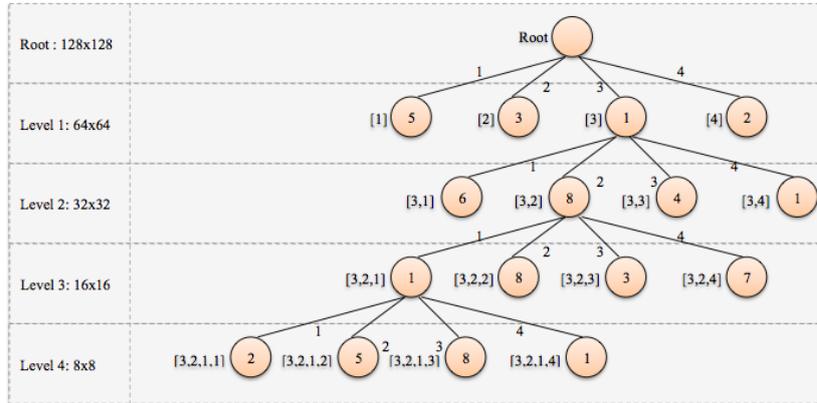
Fig. 7. The example of tree construction.

of σ the greater the number of frequent subgraphs that will be discovered. The selected value for σ will therefore influence the effectiveness of the final classifier.

Once a set of frequently occurring subgraphs has been identified these can be arranged into a feature vector representation such that each vector element indicates the presence or absence of a particular subgraph within respect to each household (record). Table 1 shows the format of the result. The rows in the table represent individual household (records) numbered from 1 to m, and the columns individual frequent subgraphs represented by the set $\{S_1, S_2, \ldots, S_n\}$. The values 0 or 1 indicate the absence or presence of the associated subgraph for the record in that row. This feaster vector representation is ideally suited to both the application of classifier generation algorithms and the future usage of the generated classifiers.

Table 1. The example of Feature Vector

Vector	S_1	S_2	S_3	S_4	S_5	...	S_n
1	1	0	1	1	1	...	1
2	1	1	0	1	1	...	0
3	1	0	1	0	1	...	1
4	0	0	1	1	0	...	0
...	...	...	...	...	...	...	...
m	0	1	1	0	1	...	1

6   Evaluation

The evaluation of the proposed population estimation mining process is presented in this section. The evaluation was conducted by considering a specific case study directed at a rural area of Ethiopia. Sub-section 6.1 provides further detail of this study

area. The overall aim of the evaluation was to provide evidence that census data can be effectively collected using the proposed approach. To this end three sets of experiments were conducted as follows:

1. A set of experiments to identify the most appropriate support threshold for use with respect to the frequent subgraph mining (Sub-section 6.2).
2. A set of experiments to analyse of the most appropriate number (k) of features to retain during feature selection (Sub-section 6.3).
3. A set of experiments to determine the most appropriate classifier generation paradigm. To this end a selection of different classifier generators, taken from the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench.[2], were considered (Sub-section 6.4).

Each is discussed in further detail in Sub-sections 6.2 to 6.4 below. Ten fold Cross-Validation (TCV) was applied throughout and performance recorded in terms of: (i) accuracy (AC), (ii) area under the ROC curve (AUC), (iii) sensitivity (SN), (iv) specificity (SP) and (v) precision (PR).

Fig. 8. Examples of satellite images from test Sites A and B.

6.1   Data Set

For the evaluation reported in this section a training dataset comprising 120 records was used: 70 records from Site A and 50 from Site B (see Figure 1). High resolution satellite images were used, obtained from GeoEye at a 50cm ground resolution, made publicly available by Google Earth[3]. The images for site A were dated 22 August 2009, while those for Site B were dated 11 February 2012. The significance is that the Site A images were captured during June to August which is the "rainy season" in Ethiopia, and thus the households tend to have a green background; while the Site B images were captured

---

[2]http://www.cs.waikato.ac.nz/ml/weka/
[3]http://www.google.com/earth/index.html

during September to February which is the "dry season", hence the images tended to have light-brown background. The contrast with respect to images obtained during the rainy season was much greater than those obtained during the dry season, hence it was conjectured that the dry season images (Site B) would provide more of a challenge.

The corresponding ground household information, required for the training data, included family size and household coordinate (latitude and longitude). This was collected by University of Liverpool ground staff in May 2011 and July 2012. With respect to all 120 records the following was noted with respect to family size: (i) the minimum was 2, (ii) the maximum was 12, (iii) the average was 6.31, (iv) the medium was 6 and (v) the stand deviation was 2.56. Therefore, for evaluation proposes, the labeled households were separated into three classes: (i) *small family* (less than 6 people), (ii) *medium family* (between 6 and 8 people), and (iii) *large family* (more than 8 people). Some statistics concerning the class distributions for the Sites A and B data sets are presented in 2.

Table 2. Class label distribution for Site A and B data sets

Location	Small family	Medium family	Large family	Total
Site A	28	32	10	70
Site B	19	21	10	50
Total	47	53	20	120

Table 3. Number of identified features produced using a range of σ values with respect to the Site A and B data

minSup	Site A	Site B
10	757	420
20	149	119
30	49	60
40	24	39
50	12	19

6.2 Subgraph Mining

In order to investigate the effect the value of the subgraph mining support threshold σ had on classification performance a sequence of different σ values were considered ranging from 10 to 50 incrementing in steps of 10. The number of features (subgraphs) generated in each case are presented in Table 3. From the table it can be seen that, as would be expected, the number of identified subgraphs decreases as the value for σ

increases (and vice-versa). Note that attempts to conduct the sun-graph mining using σ values of less than 10 proved unsuccessful due to the computational resource required (subgraph mining is computationally expensive).

To reduce the overall size of the feature space Information Gain feature selection was applied to select the top k features ($k = 25$ was used for the experiments reported in Sub-section 6.3, had revealed that this was the most appropriate value for k). Naive Bayes classification was applied with respect to each of the resulting datasets (see Sub-section 6.4 is the best overall result). The results are presented in Table 4 (where best values are highlighted in bold font). From the table it can be observed that best results were obtained using $\sigma = 10$ for both Site A (rainy season) and Site B (dry season); sensitivity values of 0.671 and 0.780, and AUC values of 0.769 and 0.829 respectively.

Table 4. Classification outcomes using a range of σ values with respect to the Site A and B data ($k = 25$)

minSup	Site A					Site B				
	AC	AUC	PR	SN	SP	AC	AUC	PR	SN	SP
10	**0.671**	**0.769**	**0.686**	**0.671**	**0.765**	**0.780**	**0.829**	**757**	**0.780**	**0.813**
20	0.571	0.660	0.582	0.571	0.741	0.580	0.771	0.579	0.580	0.768
30	0.443	0.565	0.459	0.443	0.670	0.540	0.698	0.544	0.540	0.749
40	0.343	0.389	0.340	0.343	0.555	0.440	0.615	0.440	0.440	0.688
50	0.357	0.426	0.320	0.357	0.538	0.340	0.459	0.385	0.340	0.615

6.3   Feature Selection

To identify the effect on classification performance of the value of k with respect to the adopted Information Gain feature selection method, a sequence of experiments was conducted using a range of values fore k from 10 to 30 incrementing in step of 5. For the experiments $\sigma = 10$ was used because previous experiments, reported in Sub-section 4, had indicated that a value of $\sigma = 10$ produced the best performance. The Naive Bayes classifier was again adopted. The results produced are presented in Table 5. From the table it can be seen that: (i) for Site A the best result tended to be obtained using $k = 25$ (sensitivity = 0.671 and AUC = 0.769), and (ii) for Site B the best result tended to be obtained using either $k = 20$ (sensitivity = 0.780 and AUC = 0.838) or $k = 25$ (sensitivity = 0.780 and AUC = 0.829). Hence we conclude $k = 25$ to be the most appropriate value for k (this is why $k = 25$ was used with respect to the experiments reported in Sub-section 6.2).

6.4   Classification Learning Methods

To determine the most appropriate classification method eight different algorithms were considered: (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) Sequential Minimal Optimisation (SMO), (vii) Logistic

Table 5. Comparison of different values of k with respect to Information Gain feature selection in terms of classification performance

Number of k	Site A					Site B				
	AC	AUC	PR	SN	SP	AC	AUC	PR	SN	SP
10	0.557	0.710	0.559	0.557	0.708	0.760	0.821	0.775	0.760	0.850
15	0.571	0.753	0.576	0.710	0.740	0.740	0.838	0.751	0.740	0.837
20	0.657	0.769	0.678	0.657	0.743	**0.780**	**0.838**	**0.786**	**0.780**	**0.859**
25	**0.671**	**0.769**	**0.686**	**0.671**	**0.765**	0.780	0.829	0.805	0.780	0.852
30	0.629	0.761	0.631	0.629	0.746	0.720	0.836	0.757	0.720	0.813

Regression and (viii) Neural Networks. For the experiments $\sigma = 10$ was used because this produced the best result with respect to the experiments reported in Sub-section 6.2, together with $k = 25$ for feature selection because this produced the best result with respect to the experiments reported inSub-section 6.3. The obtained results are presented in Table 6. From the Table it can be observed that:

– With respect to the Site A data, the best results (Sensitivity = 0.700 and AUC = 0.718) were obtained using the AODE classifier, and with respect to the Site B data, the best results (Sensitivity = 0.780 and AUC = 0.829) were obtained using the Naive Bayes classifier .
– The C4.5 and Logistic regression classifiers did not perform well for the Site A.
– The Logistic Regression and Neural Network classifiers did not perform well for the Site B.

Thus, in conclusion a number of different classifiers produced a good performance, but overall the Naive Bayes classifier proved to be the most effective.

Table 6. Comparison of different classifier generators in terms of classification performance

Learning method	Site A					Site B				
	AC	AUC	PR	SN	SP	AC	AUC	PR	SN	SP
C4.5	0.529	0.620	0.526	0.529	0.691	0.600	0.711	0.605	0.600	0.769
Naive Bayes	0.671	0.769	0.686	0.671	0.765	**0.780**	**0.829**	**0.805**	**0.780**	**0.852**
AODE	**0.700**	**0.809**	**0.713**	**0.700**	**0.779**	0.740	0.820	0.738	0.740	0.842
Bayes Network	0.657	0.775	0.672	0.657	0.762	0.760	0.823	0.767	0.760	0.847
RBF Network	0.571	0.709	0.579	0.571	0.707	0.680	0.750	0.668	0.680	0.820
SMO	0.557	0.663	0.557	0.557	0.703	0.620	0.737	0.618	0.620	0.778
Logistic Regression	0.500	0.659	0.502	0.500	0.674	0.540	0.635	0.543	0.540	0.739
Neural Network	0.686	0.774	0.693	0.686	0.979	0.580	0.691	0.584	0.580	0.771

7   Conclusion

In this paper a framework for population estimation mining (census mining) was proposed founded on the concept of applying classification techniques to satellite imagery.

Of particular note is the subgraph feature vector representation that was used to encode household imagery. The proposed framework was evaluated using test data collected from two villages in the Ethiopian hinterland. The conducted evaluation indicated that when using a minimum support threshold of $\sigma = 10$ for the subgraph mining, a value of $k = 25$ for feature selection and either a Naive Bayes, good results could be obtained. For future work the research team intend to conduct a large scale census collection exercise using the proposed framework.

References

1. A.S. Alsalman and A.E. Ali. Population Estimation From High Resolution Satellite Imagery: A Case Study From Khartoum. *Emirates Journal for Engineering Research*, 16(1):63–69, 2011.
2. 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 *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, pages 691–694, 2007.
3. R.A. Cole, D.G. NovickNovick, D. Burnett, B. Hansen, S. Sutton, and M. Fant. Towards automatic collection of the us census. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume i, pages I/93–I/96 vol.1, 1994.
4. 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.
5. R. C. Gonzalez and R. E. Woods. *Digital Image Processing (3rd Edition)*. Pearson Prentice Hall, 3 edition, 2007.
6. J. Han and K. Ma. Fuzzy color histogram and its use in color image retrieval. *Image Processing, IEEE Transactions on*, 11(8):944–952, 2002.
7. Y. Javed, M.M. Khan, and J. Chanussot. Population density estimation using textons. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 2206–2209, 2012.
8. 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(13):44–66, 2004. ¡ce:title¿Special Issue: Colour for Image Indexing and Retrieval¡/ce:title¿.
9. R. J. Schalkoff. *Digital Image Processing and Computer Vision*. Wiley, 1989.
10. A. Vijayaraj and P. DineshKumar. Article:design and implementation of census data collection system using pda. *International Journal of Computer Applications*, 9(9):28–32, November 2010. Published By Foundation of Computer Science.
11. S.L. Wang and A. Liew. Information-based color feature representation for image classification. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 6, pages VI – 353–VI – 356, 2007.
12. S. Wu, X. Qiu, and L. Wang. Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience and Remote Sensing*, 42(1):58–74, January 2005.
13. X. Yan and Jiawei Han. gspan: graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 721–724, 2002.
14. Y. Zhang, R. He, and M. Jian. Comparison of two methods for texture image classification. In *Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on*, volume 1, pages 65–68, 2009.