Second Year Progress Report: Obtaining Census Data by Applying Image Classification Techniques to Satellite Imagery

Kwankamon Dittakan
Department of Computer Science, University of Liverpool

31 May 2013

1 Introduction

The collection of census data is an important source of statistical information to support decision-making processes. However, traditional census collection approaches are both expensive and time consuming. This is especially the case in rural area where the communication and transportation infrastructure is not as robust as in urban areas. The analysis of high-resolution satellite imagery provides a useful alternative way to collect census data, which is significantly cheaper than traditional methods (but may be less accuracy). The idea is to build a classifier that can label satellite images of households according to Family size which can then be used to collect census information. The key challenge here is the nature of the image representation required to support the effective generation of an appropriate classifier. Three different representations have been considered so far: (i) Colour histograms, (ii) Local Binary Patterns and (ii) frequent sub-graphs. Each represented household in the training data had a Family size class label associated with it. This data was then used to produce a number of classifiers, using the three different representations, which could then be evaluated by using them to predict household sizes.

The intention of the research is thus to investigate and evaluate algorithms and processes that can be used to build the classification model and produce the desired census data. To act as a focus for the work two rural areas in Ethiopia, for which details concerning individual households have been collected by University of Liverpool field staff, have been used. These households can be isolated using image segmentation techniques. Each household can then be represented using one of the techniques identified to date. The question to be addressed by the research is thus:

Can effective census data be obtained by applying image classification techniques to satellite imagery?

This research question encompass a number of issues as follows:

1. How to segment the satellite images so that we can identify each individual household and store this information as a sub-image?
2. What is the information that should be extracted from the identified sub-images and how can this information best be extracted?
3. Once the desired information has been extracted what is the best way of representing these images so as to support the effective generation and usage of classifiers?
4. What are the most appropriate classification techniques for predicting census data from given data in the context of (1)-(3)?
5. How to deal with "overlapping" satellite images?

The aim of this report is to describe the activities that have been done in the first and second year of the PhD programme of research. A description of work completed so far is given in Section 2, followed by section 3 which describes the future work to be completed during the following year.A work plan is presented in Section 4. Finally, Section 5 shows the published work to date.

2 Work to date

This section describes the work that has been conducted to date, namely: (i) the proposed Census mining framework (Sub-section 2.1) (ii) satellite image segmentation (Sub-section 2.2), (iii) the three image representation techniques developed so far (Sub-section 2.3) and (iv) experiments and Evaluations (Sub-section 2.4).

2.1 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 1. The framework comprised two phases (as represented by the rectangular boxes): (i) Preprocessing and (ii) Classification.

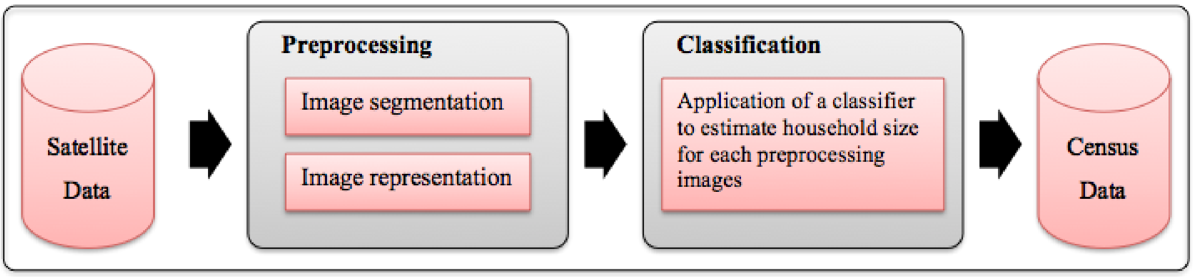

Figure 1: Proposed census mining from satellite imagery framework

During the preprocessing phase the input data is collated and prepared. The required preprocessing consists of two steps: image segmentation and image representation. Given a satellite image of a prescribed area this is segmented so as to identify a set of individual households. Each identified household image is translated into some appropriate representation that allows for application of a classifier generator. In the classification phase, after the households have been segmented and appropriately represented, the classification can take place.

2.2 Image Segmentation

This section presents a brief overview of the image segmentation process as applied to the input data. This segmentation process was described in detail in [1]. The image segmentation comprises three individual stages: (i) coarse segmentation, (ii) image enhancement and (iii) fine segmentation. In the first stage (coarse segmentation) the input satellite imagery is roughly separated into a set of sub-images covering (typically) between one and four households each depending on the nature of the image. 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.

2.3 Image 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 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. More specifically the idea is to segment satellite images so as to obtain pixel collections describing individual households and represent these collections using an appropriate representation to which a classifier generator can be applied. As noted above, three potential representations were considered: Colour Histograms, Local Binary Pattern (LBP) and Graph-based structure formats. Evaluation was conducted using information on households obtained from a ground truth survey linked to appropriate satellite images.

(i) Colour 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 [2, 3, 7] which one set is for 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 [6]. 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.

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 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. Thus on completion of the histogram based representation stage each household is represented using a feature vector of length 237 ($224 + 13 = 237$).

(ii) Local Binary Pattern

Local Binary Patterns (LBPs) are generally used for representing image texture [4]. 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 (radius) 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. For evaluation purposes three different variations of the LBP representation were used: 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. The resulting LBP representation were conceptualised in terms of a 2^R dimensional feature vector where each element represented a possible LBP value. 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). Thus on completion of the LBP based representation stage each household is represented using a feature vector of length 266 ($2^R + 10 = 2^8 + 10 = 256 + 10 = 266$).

(iii) Graph-based Structure

Quadtrees have been used extensively in the context of image processing (see for example [5]). However, the quadtree representation does not lend itself to ready incorporation in 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. The graph-based representation process consists four steps: (i) quadtree decomposition, (ii) tree construction, (iii) frequent subgraph mining and (iv) feature vector transformation.

The first step, 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", until either: (i) uniform tiles (quadrants) were arrived at or (ii) a maximum level of decomposition was reached. The generated decomposition was then stored in a quad tree format. The nodes 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. 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.

The quadtree (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 [8] was adopted. This uses the concept of a support threshold σ to define the concept of a frequent subgraph, the lower the value 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. The number of features (subgraphs) generated in each case are presented in Table 1. 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).

Table 1: 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

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 with respect to each household (record). 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 input to classifier generation algorithms and the future usage of the generated classifiers.

2.4 Experiments and Evaluations

To evaluate the proposed colour histogram, LBP and graph-baseds representations in the context of classifier generation for census data collection, the labeled training data set that was used was collected from a rural area in Ethiopia, including family size, latitude and longitude. Thus, for evaluation purposes the households ware divided into three classes: (i) *small family* (less then 6 people), (ii) *medium family* (between 6 and 8 people), and (iii) *large family* (more than 8 people). The corresponding satellite imagery was obtain from GeoEye at a 50cm ground resolution, mad publicly available by Google Earth. The images for site A were dated 22 August 2009, while those for Site B were dated 11 February 2012.

The preprocessing phase was applied to the satellite imagey (image segmentation and image representation) as described above. The resulting feature space was then discretization so that each continuously valued attribute was converted into a set of ranged attributes. Then a number of alternative classification learning methods were applied. The Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench was used for classifier generation purposes. Ten-fold Cross Validation (TCV) was used throughout. The performances of each classifier was recorded using accuracy, sensitivity, specificity, precision, and the Area Under the ROC (AUC).

The experimental results presented in this report were directed at four goals. The first was to determine the effect of classification performance with respect to the different image representations considered. The second was to compare the effect of the application of various feature selection algorithms. The third was directed at an analysis of the effect that the number of selected attributes had on performance. The fourth was directed at determining the effect on classification performance of various learning methods. The main findings from these experiments were as follows:

- *The effect on classification performance using different feature selection methods*
  Five feature selection techniques were considered: (i) Chi-Squared, (ii) Gain Ratio, (iii) Information Gain, (iv) One-R and (v) Relief-F. For each data set, three variations were considered (colour histogram, LBP(8,1) and graph-based with $\sigma = 10$). Six different K values were considered and the neural network learning methods adopted. The results indicated that Information Gain provided slightly better results than Chi-Squared and Gain Ration; One-R and Relief-F produced the lowest results.

- *The effect on classification performance on different data sets*
  The three proposed image representation techniques (plus variations) were considered. For the colour histogram representation, three data sets were produced: colour histogram (CH), colour statistics (CS) and combine (CH+CS). For the LBP reprsentation seven data sets were produced: LBP(8,1), LBP(8,2), LBP(8,3), texture statistics (TS), combine (LBP(8,1) + TS), combine (LBP(8,2) + TS) and combine (LBP(8,3) + TS). For the graph-based representation, using different minimum support threshold σ in the frequent subgraph mining process ($\sigma = 10$, 20, 30, 40 and 50), five data sets were produced. For each data set, six different K values with information Gain feature selection were considered and the neural network learning methods again applied. The results indicated that with respect to the colour histogram based representation best results were obtained using CH. With respect to the LBP representation best results were producted using LBP(8,1). And with respect to the graph-based structure representation best results were produced with $\sigma = 10$. Comparing the results obtained using all three representation, LBP(8,1) produced the best overall results.

- *The effect on classification performance using different K values*
  In order to investigated the effect on classification performance using different K values ($K = 10$, 15, 20, 25, 30 and 35) three different datasets were used (colour histogram, LBP(8,1) and graph-based with $\sigma = 10$). The Information Gain feature selection process and neural network learning methods, used for the previous experiments were again adopted. The results demonstrated that $K = 20$ produced the best overall result.

- *The effect on classification performance using different learning methods*
  The nine learning methods used previously were considered with respect to the experiments directed at identifying the best learning method. The nine learning methods considered were: (i) Decision Tree (C4.5), (ii) Nave Bayes, (iii) Averaged One-Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimization (SMO), (vii)Neural Network and (viii) Logistic Regression. In each case three different dataset including colour histogram, LBP(8,1) and graph-based with σ = 10 were used. K= 20 was used together with Information Gain feature selection. The results indicated that the neural network learning algorithm produced the best results with respect to CH and LBP(8,1). While with respect to the graph-based structure representation the Naive Bayes, AODE and Bayesian Network learning algorithms produced the best results. Comparing the results obtained using all three representation, the Neural network learning algorithm produced the best overall results.

3 Future Work

A description of the work to be completed during the coming year is presented in this section. This may be itemise as follow:

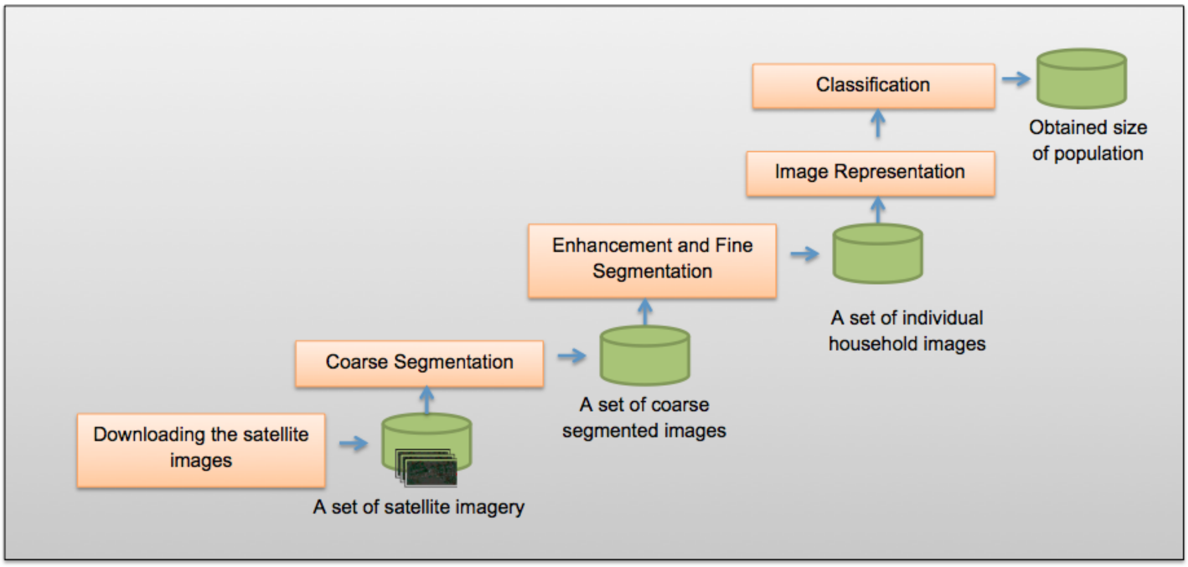
Classification
Obtained size of population
Image Representation
Enhancement and Fine Segmentation
A set of individual household images
Coarse Segmentation
A set of coarse segmented images
Downloading the satellite images
A set of satellite imagery

Figure 2: Schematic of the large scale study

- *Large scale study*
  To obtain the size of a population in a given large scale area using satellite imagery. Figure 2 presents a schematic outlining the envisioned large scale study process. This comprises five steps: (i) downloading the satellite images, (ii) Coarse segmentation, (iii) Enhancement and fine segmentation, (iv) image representation and (v) classification. Giving an initial latitude and longitude the required satellite imagery will be automatically identified and downloaded. In the coarse segmentation step, the downloaded satellite imagery is roughly segmented to give a set of large scale sub-images each covering an area of between one and four households. In the next step various image enhancement processes are applied to enhance the households identified during the coarse segmentation and then the fine segmentation process is applied to give a set of individual household images. After the households have been fine segmented they will be translated into the proposed representations. Finally, classifiers will be generated and used to classify the represented household images so that family sizes can be obtained and a census collected.

- *Overlapping problem*
  When satellite images are downloaded, the “overlapping problem” may occur. The intention is to

investigate techniques to address this problem and to generation a general purpose solution.

- *Thesis writing*
  Thesis writing using technical reports and research papers generated in previous works as the foundation for the final thesis.

4 Programme of work

A seven phases program of work was envisaged, with each phase comprising a series of “Work Packages”, as presented in following table.

Phase 1: Background work and Literature review				
WP	Start Date	End Date	Description	Deliverables
1.1	1 Oct 11	15 Oct 11	Preparing Work Plan	Programme of work (**Completed**)
1.2	16 Oct 11	10 Nov11	Review of Image Processing Techniques	Working Document 1 (**Completed**)
1.3	11 Nov 11	30 Nov 11	Review of data mining techniques (and image classification)	Working Document 2 (**Completed**)
1.4	1 Dec 11	31 Dec 11	Production of literature review (Note that updating of the literature review will continue throughout the programme of work)	Literature Review Document (**Completed**)

Phase 2: Review of data set and potential interpretation techniques				
WP	Start Date	End Date	Description	Deliverables
2.1	1 Oct 11	31 Jan 12	Preparing data (investigation of potential image enhancement, segmentation and registration techniques)	Training and test data collections (**Completed** but may need more as research progresses)

Phase 3: Census data from satellite imagery using a colour histogram format				
WP	Start Date	End Date	Description	Deliverables
3.1	1 Feb 12	15 Mar 12	Investigate techniques for converting images into a colour histogram format (or similar representation)	Technical report for applying classification technique for the colour histogram format (**Completed**)
3.2	16 Mar 12	31 May 12	Experiment using classification techniques with the colour histogram format	(**Completed**)
3.3	1 Jun 12	30 Jun 12	Evaluation of proposed technique	Research paper “Towards The Collection of Census data from Satellite Imagery using Data Mining: A Study With Respect to the Ethiopian Hinterland”, presented for refereeing to AI12 (**Completed**).

Phase 4: Census data from satellite imagery using a Tabular format				
WP	Start Date	End Date	Description	Deliverables
4.1	1 Jun 12	15 Aug 12	Investigate techniques for converting images into a tabular format. (Local Binary Pattern image representation was applied)	Technical report for applying classification technique for the tabular format (**Completed**)
4.2	16 Aug 12	15 Sep 12	Experiment using classification techniques with the tabular format	(**Completed**)
4.3	16 Sep 12	31 Oct 12	Evaluation of proposed technique	Research paper "Satellite Image Mining for Census Collection: A comparative Study with Respect to the Ethiopian Hinterland", presented for refereeing to MLDM'13 (**Completed**)

Phase 5: Census data from satellite imagery using a Quadtree format				
WP	Start Date	End Date	Description	Deliverables
5.1	1 Nov 12	15 Dec 12	Investigate techniques for converting images into a quadtrees format	Technical report for applying classification technique for the quadtree format. (**Completed**)
5.2	16 Dec 13	28 Feb 13	Experiment using graph-based classification techniques with using the quadtree format	(**Completed**)
5.3	1 Mar 13	30 Apr 13	Evaluation of proposed technique	Research paper "Population Estimation Mining Using Satellite Imagery", presented for referring to DaWak'13 (**Completed**)

Phase 6: Final Evaluation for "large scale study" and "overlap problem resolution				
6.1	1 May 13	31 July 13	Investigate techniques for large scale study and overlap problem and generation of a general process to support census collection founded on image classification	Technical report describing overlap problems resolution and general census collection process.
6.2	1 Aug 13	15 Sep 13	Experiment and evaluate proposed techniques to deal with large scale study and overlap problem	
6.3	16 Sep 13	31 Oct 13	Final evaluation and refinement of work	Final evaluation Document and research paper

Phase 7 Thesis Writing				
7.1	1 Nov 13	30 Sep 14	Thesis writing using technical reports and research papers generated in earlier sections as the foundation for the final thesis	Thesis

5 Published Work to Date

- Published papers:

1. Dittakan, K., Coenen, F. and Christley, R., *Towards The Collection of Census Data From Satellite Imagery Using Data Mining: A Syudy With Respect to the Ethiopian Hinterland.*, Proc. AI 2012. Springer, pp405-418.

- Accepted for publication:

  2. Dittakan, K., Coenen, F. and Christley, R., *Satellite Image Mining for Census Collection: A Comparative Study With Respect to the Ethiopian Hinterland*, Accepted for publication at MLDM'13.

  3. Dittakan, K., Coenen, F., Christley, R. and Wardeh, M., *Population Estimation Mining Using Satellite Imagery.*, Accepted for publication at DaWaK'13.

- In preparation:

  3. Dittakan, K., Coenen, F., Christley, R. and Wardeh, M., *Population Estimation Mining Using Satellite Imagery: A Comparative Study With Respect to the Ethiopian Hinterland*

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] R. J. Schalkoff. *Digital Image Processing and Computer Vision.* Wiley, 1989.

[6] 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.

[7] 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.

[8] 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.