Integrated Analysis of Ground Level and Aerial Image Data

Sambit Bhattacharya[1], Anil M. Cheriyadat[2]
[1]Department of Mathematics and Computer Science, Fayetteville State University, 1200
Murchison Road, USA, NC 28301, sbhattac@uncfsu.edu
[2]Geographic Information Science and Technology Group, Oak Ridge National Laboratory,
P.O. Box 2008, USA, TN 37831, cheriyadatam@ornl.gov

Abstract. Integrated analysis of ground level images with aerial image data is a new area of computer vision research which is gaining traction due to the availability of high volumes of crowd-sourced and open data that are being collected worldwide. Novel problems such as geo-location from a single or a sequence of ground level images and use of geospatial context in object recognition are active research topics. Geo-tag metadata from image data as found in social media can be utilized to extract aerial imagery surrounding the location where an image was acquired. While images found in social media repositories provide close up details of structures from ground view, aerial images have other desirable properties such as known resolution, geometry etc. The approach adopted in this project leverages features from both ground level and aerial imagery that are similarly geo-located to improve the classification of visual scenery around that place. Image feature extraction and machine learning methods are being evaluated to test the efficacy of this approach in Volunteered Geographic Information. This work can improve scenery classification for use in large scale image indexing and enhance the use of crowd-sourcing as affordable and practical means of filling information gaps in land use information.

1 INTRODUCTION

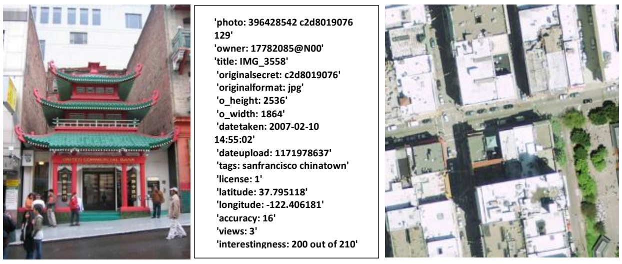
Progress in scientific understanding is being increasingly driven by our ability to collect and analyze vast repositories of data about diverse phenomena. Information capturing mobile devices, cameras etc. are now ubiquitous and growing numbers of people around the world are contributing to rapidly expanding repositories commonly described as social media. Twitter (text data), Flickr (image data), YouTube (video data) etc. are examples of such repositories which are experiencing explosive growth and are being increasingly exploited for various applications, such as situational awareness and intelligence gathering [1, 2]. Researchers are actively developing tools for social media analysis which take into account unique challenges that are not typically present in big data collected from physical phenomena. Such challenges arise due to lack of structure and missing, ambiguous or erroneous data. Analysis of such crowd-sourced data has found important applications in understanding phenomena at a global scale where geospatial patterns are of interest. In most of these applications systematic data collection using professional surveyors is prohibitively expensive so crowd-sourcing is seen as a viable option.

1.1. Related Work

The proposed research is motivated by the emerging applications of the analysis of crowd-sourced image data to the discovery of geospatial patterns. Traditional computer vision has addressed problems without reference to the geo-tag of imagery or in other words the latitude and longitude of the place where the image was captured. Recent work is realizing the value of using such Volunteered Geographic Information (VGI) in creating novel solutions to open problems and also in defining new areas of research [3]. The difficult problem of geolocalization from image content is a new area where [4] originally proposed using a data driven approach to estimate where on the surface of the earth an image was captured by clustering the best scene based matches to a large image database. Although VGI is a rich source of data there are some uncertainties regarding the quality of information that can be gained. Current research [5, 6] has thus investigated its usefulness in improving land cover classification, methods of filtering out non-informative images and also the effect of the photographers intent when capturing the image. Additionally [6] proposes using existing geographic information as training data in a weakly supervised manner. Research is addressing questions at geographic scales of space and time. A geo-wiki is proposed in [7] to improve land cover classification; [8] proposes to mine geo-tagged images for discovering cultural differences and [9] combines traces of Global Positioning System (GPS) data with image sequences for event classification.

1.2. Questions Addressed

The present work is an effort to extend the boundary of what has been so far deemed possible through the content of geo-tagged imagery alone. The goal is to find improved techniques of combining aerial imagery from the geo-tagged locations of ground level data so that a different perspective can be brought into the analysis. The goal of the analysis is to improve automated machine understanding of both aerial and ground where each data source benefits from association with the other. Figure 1 shows an example of a Flickr image with extracted metadata and associated aerial imagery. This work can result in improvement of scenery classification for use in large scale image indexing; better accuracy and extent of global land use information and enhanced use of crowd-sourcing as an affordable and practical means of filling in information gaps and verifying existing land use information, images from disaster locations etc. The objectives of this work are to answer questions that will help accomplish the goal. Firstly, what are best combinations of low level and semantic features and that extracted from ground and aerial imagery to improve their combined analysis? Here work is addressed at the open question of whether existing and frequently used object detection methods in computer vision are well suited for extracting geographic information from geo-referenced imagery [10]. Another related question is which features are best with respect to the



```
'photo: 396428542 c2d8019076
129'
'owner: 17782085@N00'
'title: IMG_3558'
 'originalsecret: c2d8019076'
 'originalformat: jpg'
 'o_height: 2536'
 'o_width: 1864'
 'datetaken: 2007-02-10
14:55:02'
 'dateupload: 1171978637'
 'tags: sanfrancisco chinatown'
 'license: 1'
 'latitude: 37.795118'
 'longitude: -122.406181'
 'accuracy: 16'
 'views: 3'
 'interestingness: 200 out of 210'
```

Figure 1 - Example of metadata extraction from ground level image (image, left) to obtain geo-tag, text descriptions etc (text box, center). The geo-tag is next used to find aerial imagery around the geo-location and finally a suitably sized aerial image patch is extracted around that location to provide aerial data (image, right).

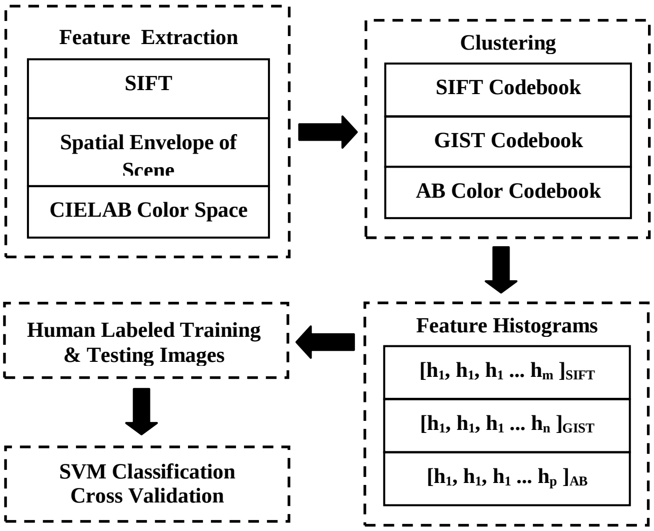
information being sought. For example humans are good at visually classifying geographical neighbourhoods as having affinities to racial and ethnic identity, income levels, occupations etc. Can machines be taught to similar classification by discovering latent features? Next, how is it possible to achieve high classification accuracy while reducing the mathematical dimensions of the features used in the analysis? This part of the work increases the efficiency of data processing and provides insights into the relevance of features. Work on characterization of aerial imagery of urban landscapes [11] shows such problems to be very high dimensional. Another interesting question is how will scene classification accuracy be impacted for ground level images with no corresponding aerial data or vice versa? Such impacts must be measured to understand the importance of the association in this analysis and also to address the missing data problems of VGI.

2 METHOD

Our approach leverages features from both ground level and aerial imagery that are similarly geo-located to improve the classification of visual scenery around that place. While images acquired from the ground may show scenery of great diversity and complexity, aerial images are also known to have similar characteristics. For instance the problems of building detection in aerial images [14] and characterization of neighbourhoods as formal or informal [11] are areas where advanced feature extraction and probabilistic modelling techniques continue to be applied for improved analysis. In this work we are investigating the use of image feature extraction from corresponding ground and aerial imagery and methods of combining them. While a vast literature exists on feature extraction [13] we have performed initial experiments on a selected few methods only with future plans to incorporate other methods. Classification using machine learning methods is performed on the representation of images as feature histograms. Currently we are concatenating feature vectors of the ground and corresponding aerial image to form a single feature vector that represents each pair. The system is summarized in a diagram in Figure 2.

2.1. Methods of Capturing Image Features

We first extract low level features for a large collection of such images – ground view images coupled with their aerial patch of pixels. Current features used are Scale Invariant Feature Transform (SIFT) [17]; spatial envelop of scene [16]; and the CIELAB colour space which is better aligned to human perception of colour [15]. We choose SIFT since compared to other techniques like corner detection it is well known to match blob-like structures seen at different scales. We expect the CIELAB colour space to capture predominant colours within geographical neighbourhoods and also the level of luminosity which may help in a classification task like indoor versus outdoor. Also the spatial envelope of a scene is known to compute numerical measures that correspond well to human concepts like naturalness, openness etc. We expect this feature method to be complementary to what the other methods provide.



Feature Extraction
SIFT
Spatial Envelope of Scene
CIELAB Color Space
Clustering
SIFT Codebook
GIST Codebook
AB Color Codebook
Feature Histograms
$[h_1, h_1, h_1 \dots h_m]_{SIFT}$
$[h_1, h_1, h_1 \dots h_n]_{GIST}$
$[h_1, h_1, h_1 \dots h_p]_{AB}$
Human Labeled Training & Testing Images
SVM Classification Cross Validation

Figure 2 - Summary of the system that transforms data from raw values to features; the features are further processed in to histograms that are estimates of underlying probability distributions of features given images classes; human labellers produce labelled image data that are divided into training and test sets; an SVM machine learning algorithm gets trained on the training set and is then tested on the test set to assess the accuracy of learning.

2.2. Representation and Classification

This project has completed the initial implementation task of developing a system under existing machine learning framework. Preliminary results have been obtained from training the system to perform a binary classification task. The results point at areas of improvement and future extensions described in the next section. The system processes data in several stages beginning with feature extraction of both ground and aerial imagery. In the next stage the features are clustered to find the centres (collectively called codebooks) that are best representative of the diversity of the data. Then the raw feature values are further processed in relation (in this case Euclidean distance) to the cluster centres. This stage of vector quantization produces feature histograms that are better suited for submitting to a machine learning program. Human labellers produce a labelled data set. In this study a binary classification of outdoor versus indoor scenery was used. A part of the labelled data is kept for training and the other part is set aside for testing. Next the labelled data are given as input to the Support Vector

Machine (SVM) classification algorithm. Best SVM parameters for an RBF kernel are selected through a grid search where are at each point of the grid (determined by selection of values for parameters) a cross-validation experiment is performed to measure accuracy. The parameters corresponding to the highest achieved accuracy are then used for SVM and the accuracy of the SVM is measured against the test data that was created earlier. Further variations on number of cluster centres (effects the quantization) and number of features sampled are used to compute accuracy levels. This helps in identifying the best choices for quantization and sample size.

3 RESULTS

To test our approach we began with a collection of around 30,000 images downloaded from Flickr that were from within a wide geographical region that covers most of the city of San Francisco, USA. The USGS Earth Explorer (http://earthexplorer.usgs.gov/) was used to download aerial imagery covering all locations (latitude and longitude were obtained from geo-tag metadata) from which these images had been acquired. A pre-processing task was to extract 512 x 512 aerial image patches centred around those geo-locations. The data thus had corresponding ground level and aerial image pairs. Next we computed the SIFT, spatial envelope and CIELAB colour features from all images. In the codebook computation stage we took random samples of values of each feature type from the computed data. We experimented with different sizes of total number of samples and observed that increasing total number of samples beyond 200K had no significant positive effect. The k-means algorithm was applied with varying number of centres as inputs – 100, 200, 400, 500 to 1000. The best results (after further computations) were achieved by starting with codebook having 400 centres. Next, around 4000 images were randomly selected and human labelled as showing either indoor or outdoor scenery. Feature histograms were computed for all images based on the codebooks. We decided to perform the binary indoor versus outdoor classification as our initial experiment. Also the human labeller used either one of the labels when the image was ambiguous such as a close up shot of face. Under these conditions the images were highly diverse, had ambiguities and the human labeller was not concerned with the accuracy of the geo-tag. From the labelled images we created a training set of around 2500 image pairs and set aside another 1500 for testing. We performed grid search with cross validation at each grid point to choose the best SVM model based on the training data. When tested the accuracy was around 70%. This was the level of accuracy when we repeated the process with only ground level images not associated with their aerial patch. In other words it pointed out that there was no significant gain by combining features from the two images in the pairs. We next manually selected a much smaller set of images (around 100) from test set based on some well known geographical features of the region where the data came from. A careful selection process was applied to pick images that came from three different types of locations – urban built-up area, open parks and facing open water body. Furthermore, images from built-up areas there were outdoor (such as parade on the road) were not included. Under these new conditions the accuracy improvement was up to 75% with the combined data over 70% using ground level image alone. The observation that can be drawn from these results is that incorporation of aerial imagery can potentially improve classification accuracy; however research has to be conduct on approaches to handle the diversity, ambiguities and inaccuracies in crowd-sourced data.

4 DISCUSSION

Future work for this project will be performed to address specific issues and also extensions to the current work. First, the quality of VGI needs to be addressed in a more systematic way since geo-tags have been observed to be non-existing or inaccurate in many instances and geo-tags can be completely unrelated to scenery e.g. a picture of the model of a landmark taken at a location far from the landmark itself. One strategy is to cluster data around the users who are uploading multiple images within short time intervals. This may lead to better understanding of user behavior and improved methods of tackling imprecision in metadata. Further experimental analysis is needed to measure effects of sample size & clustering parameters for combinations of ground level and aerial feature vectors and if alternative and better methods exist to the currently used concatenation of feature vectors. More classes of scenery will be incorporated and a method of dividing geographic space into class-homogenous regions will be developed. Statistical measures will be used to assess the quality of classifications. This is needed when available data is unbalanced as in the case of VGI.

Future extensions of this work will consider the relationship of geographically informative objects with land use classes. As an example consider ground level imagery showing green grass covered landscape that are roughly from two geographic locations. Based purely on land cover information and aerial imagery they may be classified into the same land use pattern. However ground level imagery from one location may reveal presence of objects (golfers versus cattle) that may lead to improved land use information (golf course versus pasture). Another extension will be the association of temporal image sequences with aerial imagery for characterizing geo-locations based on activity patterns.

5 CONCLUSION

This work investigates open problems in image analysis that combines data from both ground level and aerial imagery which are captured around the same geographical location. The goal is to improve scene classification. The possible uses are validation of land use and improved image indexing. Initial work has provided preliminary results based on existing approaches from computer vision and machine learning and has pointed out specific areas of work and future extensions.

6 ACKNOWLEDGEMENTS

S. Bhattacharya (author) thanks the leading scientists of the Geographic Information Science and Technology Group at Oak Ridge National Laboratory. He also acknowledges the Center for Defense and Homeland Security at Fayetteville State University.

7 REFERENCES

[1] Virtual Social Media Working Group, D.F.R.G., *Community Engagement and Social Media Best Practices*, 2013, U.S. Department of Homeland Security.

[2] Virtual Social Media Working Group, D.F.R.G., *Lessons Learned: Social Media and Hurricane Sandy*, 2013, U.S. Department of Homeland Security.

[3] Luo, J., et al., *Geotagging in multimedia and computer vision—a survey*. Multimedia Tools and Applications, 2011. **51**(1): p. 187-211.

[4] Hays, J. and A.A. Efros. IM2GPS: estimating geographic information from a single image. in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. 2008. IEEE.

[5] Estima, J. and M. Painho, Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover, in Computational Science and Its Applications–ICCSA 2013. 2013, Springer. p. 205-220.

[6] Leung, D. and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010. IEEE.

[7] Fritz, S., et al., *Geo-Wiki: An online platform for improving global land cover*. Environmental Modelling & Software, 2012. **31**: p. 110-123.

[8] Yanai, K. and B. Qiu. Mining cultural differences from a large number of geotagged photos. in Proceedings of the 18th international conference on World wide web. 2009. ACM.

[9] Yuan, J., et al. Mining GPS traces and visual words for event classification. in Proceedings of the 1st ACM international conference on Multimedia information retrieval. 2008. ACM.

[10] Leung, D. and S. Newsam. Can off-the-shelf object detectors be used to extract geographic information from geo-referenced social multimedia? in Proceedings of the 5th International Workshop on Location-Based Social Networks. 2012. ACM.

[11] Graesser, J., et al., *Image based characterization of formal and informal neighborhoods in an urban landscape*. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 2012. **5**(4): p. 1164-1176.

[12] Wang, X. and E. Grimson. Spatial latent dirichlet allocation. in Advances in Neural Information Processing Systems. 2007.

[13] Tuytelaars, T. and K. Mikolajczyk, *Local invariant feature detectors: a survey*. Foundations and Trends® in Computer Graphics and Vision, 2008. 3(3): p. 177-280.

[14] Sirmacek, B. and C. Unsalan, *A probabilistic framework to detect buildings in aerial and satellite images*. Geoscience and Remote Sensing, IEEE Transactions on, 2011. 49(1): p. 211-221.

[15] Connolly, C. and T. Fleiss, *A study of efficiency and accuracy in the transformation from RGB to CIELAB color space*. Image Processing, IEEE Transactions on, 1997. 6(7): p. 1046-1048.

[16] Oliva, A. and A. Torralba, *Modeling the shape of the scene: A holistic representation of the spatial envelope*. International journal of computer vision, 2001. 42(3): p. 145-175.

[17] Lowe, D.G. *Object recognition from local scale-invariant features*. in Computer vision, 1999. The proceedings of the seventh IEEE international conference on. 1999. Ieee.