Computational Techniques for Crop Disease Monitoring in the Developing World

John Quinn

Department of Computer Science, Makerere University
PO Box 7062, Kampala Uganda

Abstract. Tracking the spread of viral crop diseases is critically important in developing countries. It is also a problem in which several data analysis techniques can be applied in order to get more reliable information more quickly and at lower cost. This paper describes some novel ways in which computer vision, spatial modelling, active learning and optimisation can be applied in this setting, based on experiences of surveying viral diseases affecting cassava and banana crops in Uganda.

1 Introduction

The problem of monitoring the spread of infectious disease among crops in developing regions is interesting in two regards. First, it is of critical practical significance, as the effects of crop disease can be devastating in areas where one of the main forms of livelihood is subsistence farming. It is therefore important to monitor the spread of crop disease, allowing the planning of interventions and early warning of famine risk. Second, it provides an example of the scope of opportunity for applying novel data analysis methods in under-resourced parts of the world.

The standard practice currently in a country such as Uganda is for teams of trained agriculturalists to be sent to visit areas of cultivation and make assessments of crop health. A combination of factors conspire to make this process expensive, untimely and inadequate, including the scarcity of suitably trained staff, the logistical difficulty of transport, and the time required to coordinate paper reports. Although computers remain a rarity in much of the developing world, smartphones are increasingly available: for example they account for 15-20% of all phones in Kenya, projected to be at 50% by the end of 2015 [2], and there are 8 million mobile internet subscribers [6] in a country with population of 41 million. Among other benefits, the prevalence of mobile computing devices and mobile internet makes it easy to collect different types of data, and in new ways such as crowdsourcing. Once data is collected electronically, this opens up opportunities to apply computational techniques which allow the process of crop disease survey in such an environment to be reinvented entirely.

We outline here three ways in which novel data analysis techniques can be used to improve the speed, accuracy and cost-efficiency of crop disease survey, using examples of cassava and banana crops in Uganda. After briefly discussing

the mobile data collection platform we have implemented for this purpose (Section 2), we describe automated diagnosis of diseases, and image-based measurement of disease symptoms (Section 3), possibilities for incorporating spatial and spatio-temporal models for mapping (Section 4), and ways in which survey resources can be used optimally by prioritising data collection at the locations that the spatial model determines to be most informative (Section 5). These methods are currently being trialled with collaborators in the Ugandan National Crop Resources Research Institute, which specialises in cassava disease, and the Kawanda Agricultural Research Institute, which specialises in banana disease.

2 Mobile data collection

We implemented a system for collecting crop disease survey information with low cost (under 100 USD) Android phones, based on the Open Data Kit [3]. This provides a convenient interface for digitising the existing forms used by surveyors, with the ability to also collect richer data including images and GPS coordinates. Data collected on this system can be plotted on a map in real-time, see for example `http://cropmonitoring.appspot.com`. Clearly there are a number of immediate benefits from simply collecting data on a phone instead of paper, in that costs are reduced since the time needed to do data-entry and print paper forms far outweighs the costs of the phones and data, and results are immediately available. It also means that the survey can be conducted without experts being required to travel to the field; images can be collected and assessed remotely. More importantly to the purposes of this discussion, however, it allows data analysis methods to be applied which have the potential to fundamentally change the way in which the survey is conducted.

3 Automated diagnosis and symptom measurement

Since the collection of survey data with mobile devices can include images of crops, removing the requirement for experts to be physically present to carry out inspection, we next focus on automating the judgements that those experts make based on images of leaves and roots. A typical national-scale survey of cassava disease in Uganda, for example, would include judgements about disease status and levels of symptoms on around 20,000 plants. The automation of judgements on this quantity of images constitutes a considerable saving of time and resources. With sufficient labelled training data this is a feasible problem for many diseases with clearly visible symptoms, and we can therefore collect data more rapidly and at lower cost. Automatic image-based diagnosis of crop diseases from leaf images is an active field [10, 11, 7], though little previous work has focused on crops grown primarily in developing countries.

Symptoms which need to be assessed for cassava include the extent of necrosis of the roots. It is also useful to count the number of whiteflies found on the leaves, as these are the vectors for multiple viruses. Figure 1 shows the ways in which we can carry out these measurements using computer vision. Assessment

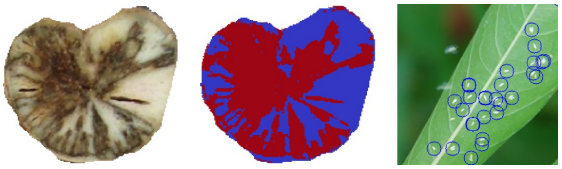

Fig. 1. Automated symptom measurement. Left: cassava root with necrotisation caused by cassava brown streak disease; center: classification of pixels to measure proportion of necrotisation; right: whitely count on cassava leaf.

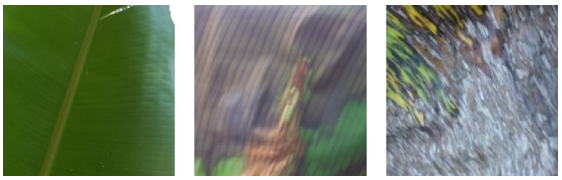

Fig. 2. Banana leaf image patches. Left: healthy leaf; center: banana bacterial wilt; right: black sitagoka disease.

of roots is currently done in the manual survey by assigning root samples to one of five categories, from completely healthy to completely necrotised. The main problem with this process is that the intermediate grades are easily confused; automating the process with image processing leads to more accurate and standardised results, removing the variability caused by different surveyors. Counting whiteflies on leaves is an infuriating and slow task for surveyors. The underside of a cassava leaf might have hundreds of these small, mobile insects, hence accurate counts are not feasible. In image processing terms, however, this is not a difficult problem, being essentially a form of blob detection.

Identification of viral diseases from leaf images is also possible given labelled data for training a classification model. Figure 2 shows examples of a healthy leaf surface and two diseases common in Uganda, banana bacterial wilt and black sitagoka disease. We have found that classification based on colour histogram features gives good results, though the incorporation of texture features is likely to improve this further. We have had similar experiences with diagnosis of cassava diseases from leaf images [1].

We have also found that with such straightforward classification techniques, it is possible to implement this process directly on the phone being used for the survey for real-time feedback. Figure 3 shows how the system works when these elements are combined. Capturing a cassava leaf image on the phone allows us to obtain an immediate diagnosis, which is uploaded to a server and plotted on a map online.

Fig. 3. Phone based survey with automated diagnosis. Left: mobile-phone based survey of cassava field; center: software on the phone detects cassava mosaic disease from leaf appearance; right: data collected with the phone is instantly uploaded to the web.

4 Incorporating a spatial model

Models of crop disease are used for understanding the spread or severity of an epidemic, predicting the future spread of infection, and choosing disease management strategies. Common to all of these problems is the notion of spatial interpolation. Observations are made at a few sample sites, and from these we infer the distribution across the entire spatial field of interest. Standard approaches to this problem (reviewed in [9]) include the use of spatial autocorrelation, or Gaussian process regression [5]. Often the extent to which each plant is affected by disease is quantified in ordinal categories, in which case a spatial model which makes efficient use of the available data is Gaussian process ordinal regression [8]. Temporal dynamics can be added to these models, allowing forecasts to be made.

4.1 Combining diagnosis and mapping

The above tasks of estimating the density of an infectious disease in space and diagnosing that disease in individual cases (as in Section 3) are generally done separately. Informally, a surveyor may be aware of outbreaks of a disease in particular places or seasonal variations in disease risk, and they may interpret test results accordingly. But the diagnosis is not usually formally coupled with estimates of disease risk from the emerging spatial model.

The tasks of mapping disease density over space and time and of diagnosing individual cases are complementary, however. A "risk map" can be used to give a prior in diagnosis of an individual plant with a known location. In turn, the results of individual diagnoses can be used to update the map in a more effective way than simply making hard decisions about infection statuses and using summary count data for the update. The potential for combining maps and diagnosis in this way comes about with the possibility of performing diagnosis with networked location aware devices that can carry out the necessary calculations, as discussed in Section 3. In practice, this combined inference of spatial disease

density and diagnosis in individual cases can be done with multi-scale Bayesian models, as described in [4]. By selecting an appropriate model structure, this can be done tractably even for very large numbers of individual plants as in the case of a national survey. This can improve both the accuracy of the risk map and of individual diagnoses, since the uncertainty in both tasks is jointly modeled.

5 Optimising survey resources

A probabilistic spatial or spatio-temporal model is useful not just in building up a picture of the disease map, but in knowing which locations would be most informative for collecting new data. While this was impossible in the traditional paper-based survey system, in which data entry would happen after the return of surveyors, the methodology described in this paper allows models to be learned in real-time as data is collected in the field. Therefore our models can be used to guide surveyors to collect more valuable data, holding fixed their budgeted number of samples.

This problem is essentially active learning, in which we prefer to collect data from locations in which the model has the lowest confidence. For example, in a Gaussian process model, we prefer to sample from locations where the density estimate has the highest covariance with the data already collected. This approach would be suitable for example in a crowd sourcing setting: if phones were given to agricultural extension workers across the country, and micro-payments are made to those workers in return for sending image data, it would be possible to adjust the levels of those payments based on location in order to use the budget optimally with respect to building an informative model.

When we attempt to direct the progress of a survey in which data collection teams are sent to travel around the country, the situation is a little different. There is a fixed travel budget, e.g. for fuel, and we cannot simply collect data from arbitrary locations on the map. Considering the constraints of being able to travel along a given road network with some budget, this optimisation problem is in general very complex. However, we can simplify this constraint somewhat by considering that in rural parts of the developing world, the road network is often sparse. This makes it reasonable to assume that survey teams will follow a set route, corresponding to a one dimensional manifold $\mathcal{R}$ within the spatial field. With a survey budget allowing k stops, we are interested in finding a set of points along $\mathcal{R}$ that maximise the informativeness of the survey. Under this constraint, optimisation is tractable with a Monte Carlo algorithm [8], where we recompute after each stop the optimal next sample location based on the spatial model given the most recent observation. This can also be done for multiple groups of surveyors simultaneously traveling along different routes.

6 Discussion

This paper has outlined various ways in which computational techniques can make crop disease survey more effective given tight resource constraints. It is an

illustration of one of the ways in which data analysis can be used to address problems in the developing world, where we often wish to automate the judgements of experts who are in short supply, collect intelligence about socio-economic or environmental conditions from different, noisy data sources, or optimise the allocation of some scarce resource. Similar methods can be directly applied to the survey and diagnosis of human disease, for example, another active area of current work.

References

1. J.R. Aduwo, E Mwebaze, and J.A. Quinn. Automated vision based diagnosis of cassava mosaic disease. In *Proceedings of the ICDM Workshop on Data Mining in Agriculture*, 2010.
2. Anonymous. The Arrival of Smartphones and the Great Scramble for Data. *The East African*, 25 May 2013.
3. C. Hartung, A. Lerer, Y. Anokwa, C. Tseng, W. Brunette, and G. Borriello. Open Data Kit: Tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, 2010.
4. M. Mubangizi, C. Ikae, A. Spiliopoulou, and J.A. Quinn. Coupling spatiotemporal disease modeling with diagnosis. In *Proceedings of the International Conference on Artificial Intelligence (AAAI)*, 2012.
5. M.R. Nelson, T.V. Orum, and R. Jaime-Garcia. Applications of geographic information systems and geostatistics in plant disease epidemiology and management. *Plant Disease*, 83:308–319, 1999.
6. Communications Commission of Kenya. Quarterly sector statistics report, July–September 2012.
7. A.J. Perez, F. Lopeza, J.V. Benlloch, and S. Christensen. Colour and shape analysis techniques for weed detection in cereal fields. *Computers and Electronics in Agriculture*, 25(3):197–212, 2000.
8. J.A. Quinn, K. Leyton-Brown, and E. Mwebaze. Modeling and monitoring crop disease in developing countries. In *Proceedings of the International Conference on Artificial Intelligence (AAAI)*, 2011.
9. A. van Maanen and X.-M. Xu. Modelling plant disease epidemics. *European Journal of Plant Pathology*, 109:669–682, 2003.
10. L. Wang, T. Yang, and Y. Tian. Crop disease leaf image segmentation method based on color features. In *Computer And Computing Technologies In Agriculture, Volume I*. Springer, 2008.
11. Z. Zhihua, X. Guo, C. Zhao, S. Lu, and W. Wen. An algorithm for segmenting spots in crop leaf disease image with complicated background. *Sensor Letters*, 8:61–65, 2010.