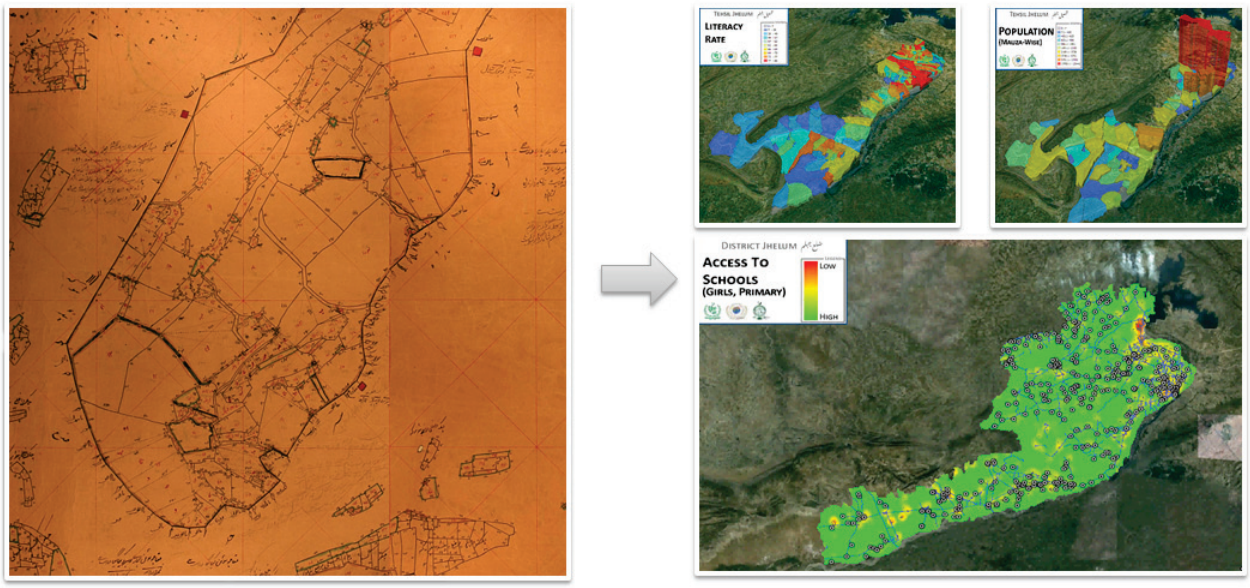
# Research Proposal for
# German-Pakistani Research Collaboration Programme




# MAPPING RURAL PAKISTAN

## FROM COLONIAL-ERA VILLAGE MAPS TO A MODERN GIS SYSTEM FOR MAPPING SOCIO-ECONOMIC AND DEMOGRAPHIC DATA

IUPR University of Kaiserslautern

**University of Kaiserslautern**
Image Understanding and Pattern Recognition (IUPR) Research
Prof. Dr. Thomas Breuel

LUMS

**Lahore University of Management Sciences (LUMS)**
Computer Vision Laboratory
Dr. Sohaib Khan

# Mapping Rural Pakistan

## From Colonial-Era Village Maps to Modern GIS Systems for Mapping Socio-Economic and Demographic Data

## 1. General Information

| | |
|---|---|
| **Project Title** | Mapping Rural Pakistan: From Colonial-Era Village Maps to Modern GIS Systems for Mapping Socio-Economic and Demographic Data |
| **Institution (Germany)** | Technical University of Kaiserslautern |
| **Supervisor (Germany)** | **Prof. Dr. Thomas M. Breuel**<br>Building. 48, Room 459<br>Technical University of Kaiserslautern<br>Gottlieb-Daimler-Str.<br>67663 Kaiserslautern, Germany<br>Ph: +49 (0)631 205-3456 Fax: +49 (0)631 205-3357<br>Email: secretary@iupr.com |
| **Institution (Pakistan)** | Lahore University of Management Sciences (LUMS) |
| **Supervisor (Pakistan)** | **Dr. Sohaib A. Khan**<br>School of Science and Engineering, Room 123-A<br>Lahore University of Management Sciences<br>Sector U, DHA<br>Lahore 54792, Pakistan<br>Ph: +92 333 431-7478<br>Email: sohaib@lums.edu.pk |
| **Other Members in Technical Team** | Dr. Faisal Shafait (Associated Researcher, DFKI, Univ of Kaiserslautern)<br>Syed Saqib Bukhari (PhD Student, IUPR, TU Kaiserslautern) |
| **Technology Area** | Computer Science, Geographical Information Systems (GIS) |
| **Track** | Document Analysis, Image Understanding, Image Processing |
| **Duration** | The project's intended duration is 3 years, but the initial duration of funding is for 2 years, subject to extension based on review |
| **Start Date** | August 2011 |
| **Proposed Budget** | € 70,000 (first two years only) |

## 2. Summary

A substantial impediment to socio-economic development, good governance and effective decision-making for rural Pakistan is the absence of administrative, demographic and thematic maps. From official products of Survey of Pakistan to crowd-sourced Google-Maps, rural Pakistan mostly is a large empty void. Detailed maps that can support decision-making are simply non-existent, thus limiting the capacity across government departments for tasks such as locating a new school, coordinating response to natural disasters or simply comparing indicators across communities.

The proposed project presents a novel and practical solution to change this state of affairs, through scalable and automatic vectorization and mosaicing of old colonial-era village (mauza) maps. These original maps, hand drawn through field surveys decades ago, are stored in district record rooms across the provinces, and contain detailed mapping of each mauza and its land parcels. The Pakistani

side has already demonstrated recently that these maps can be mosaiced and geo-referenced with satellite imagery – creating for the first time an entire district from 5000 photographed maps, by painstakingly piecing them together like a giant jigsaw puzzle.

However, the current approach requires laborious manual work. With an estimated 120,000+ map sheets containing more than 40 million land parcels in just the Punjab province, it is inconceivable that this task can be completed manually. We propose to extract land parcel boundaries, grid lines and numeric identifiers through image processing techniques, hence automating the entire process and making it scalable for large areas.

The project brings together synergistic strengths of both research groups. The Pakistani side has completed a preliminary study, demonstrating the effectiveness and the utility of the mapping methodology and has identified the bottlenecks. The German research group is world renowned in using image processing for automated document analysis, with special expertise on digitizing archival documents photographed through low-cost digital cameras.

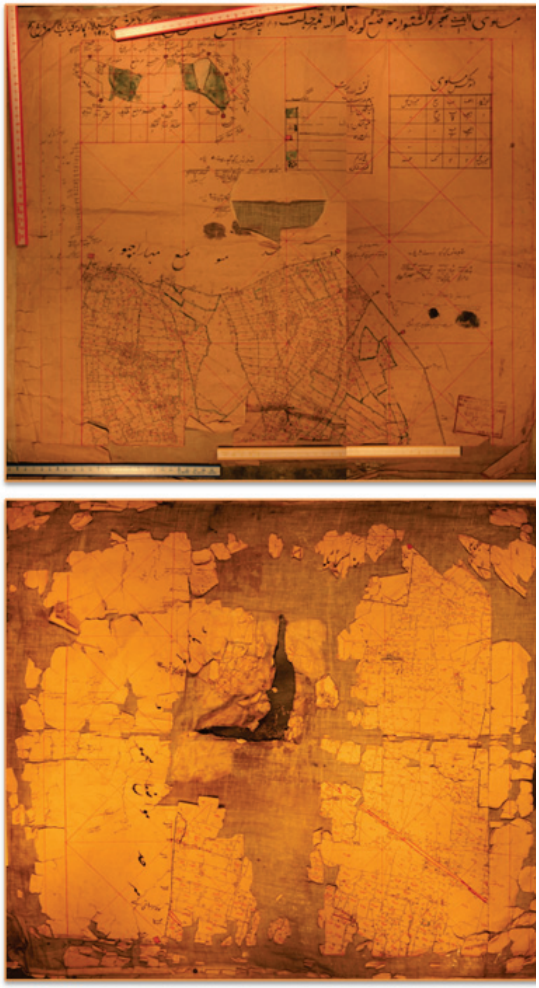


**Figure 1: [Top] Scan of a typical *masavi* (British-era village map), which is torn at a few places. [Bottom] Scan of a heavily damaged *masavi***

The project has immense practical significance, as the Punjab government has already started digitizing these maps across the province, yet it presents significant scientific challenges due to the age of the maps and their varied state of preservation. It therefore offers a unique scientific collaboration opportunity with significant socio-economic impact.

## 3. Introduction

The dearth of socio-economic and demographic mapping in rural Pakistan presents a significant barrier to good governance, effective decision making, socio-economic development and informed analysis. Detailed maps are simply unavailable for most applications, and even rudimentary mapping solutions, such as that of administrative boundaries or census data, do not exist. The corresponding lack of decision-support tools, say, for example, for the problem of identifying the location of a new primary school, forces blind and arbitrary choices that lack transparency and may easily be swayed by personal preferences or vested interests. Moreover, citizen who were to receive those services have little information about their developmental rights, entitlements and comparative information of other similar communities.

The key bottleneck that has led to this state of affairs is not the lack of appropriate datasets, but the inability to map them. Most critical datasets, such as population census, agricultural census and education data contain attributes of each *mauza* (revenue village) which is the smallest administrative

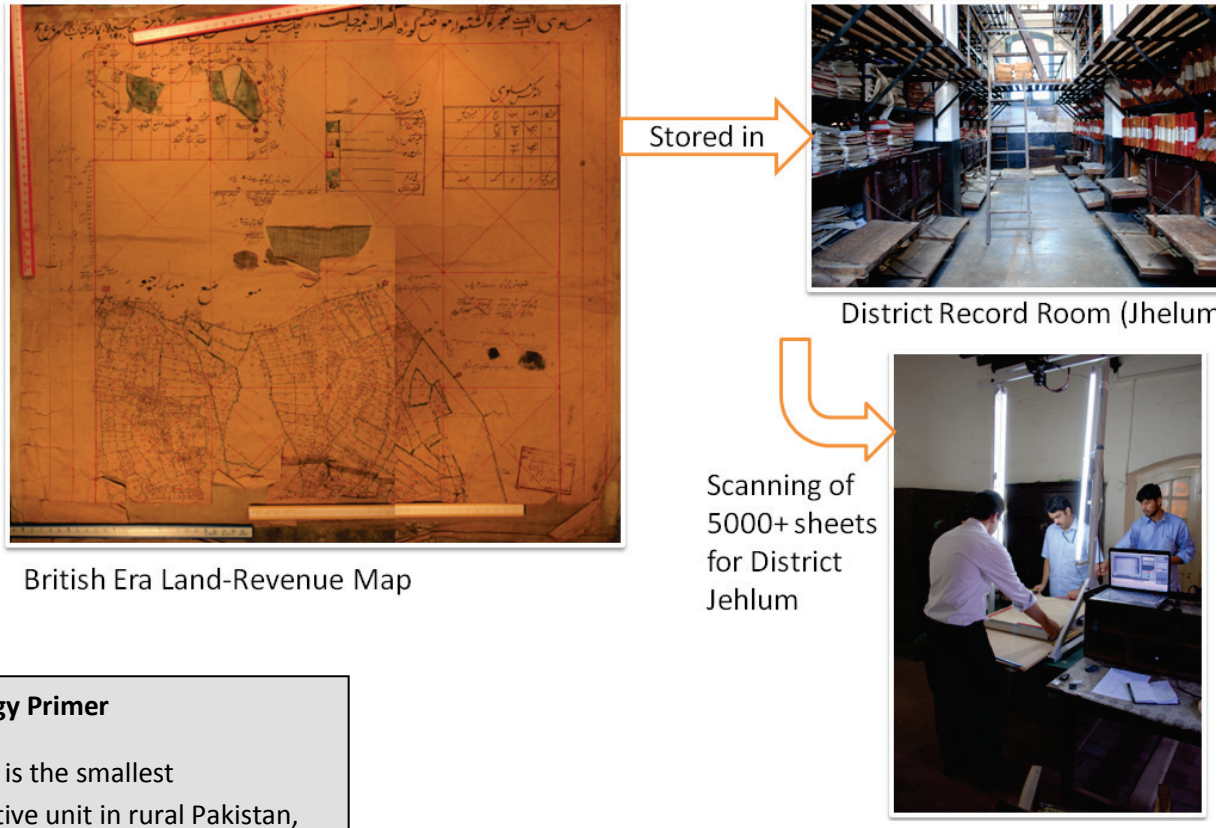



**Figure 2: Scanning operation of more than 5000 sheets of British Era Mauza Maps.**

**Terminology Primer**

The *mauza* is the smallest administrative unit in rural Pakistan, also termed as revenue village. Punjab, Pakistan's most populous province, contains 26,000 mauzas, grouped in 36 districts. A mauza is not merely the settled area of a village, but also encompasses adjoining agricultural land and grazing pastures. The average area of a mauza in Punjab is 7 $km^2$ though their size variation can be substantial.

For land record purposes, mauza maps are drawn on 30in x 24in sheets, called *masavi*, typically at the scale of 1in : 220ft. Only two copies of each map exist. These maps form the basis of all land transactions. In District Jhelum, the most recent maps were drawn in 1940.

A mauza is divided into land parcels called *khasra*. A khasra is an individual field in the village. Each khasra has a unique numeric identifier. We estimate more than 40 million khasras in Punjab.

unit in rural Pakistan. However, no integrated and geo-referenced mauza map exists, thus making it impossible to visualize these critical socio-economic datasets.

The role of having detailed administrative boundaries is critical to any mapping solution. The *mauza* is the smallest administrative unit and forms the basis of all administrative boundaries at rural level. The significance of these boundaries is immense. Most socio-economic and demographic data at the rural level, such as census data, national identity cards and patients reporting at a hospital, contain the mauza name. Yet these datasets currently are not geo-referenced, existing in tabular forms and not available on a map. Current maps of Survey of Pakistan (Pakistan's national mapping agency) show administrative boundaries only till *tehsil* (sub-district) level. To understand the difference in coarseness, it may be mentioned that there are about 120 *tehsils* in Punjab but more than 26,000 *mauzas*. By building a *mauza*-level map of District Jhelum in our recent work, we were able to show the census data on a geo-referenced map for the first time, potentially enabling inference of spatial trends for more effective decision-making.

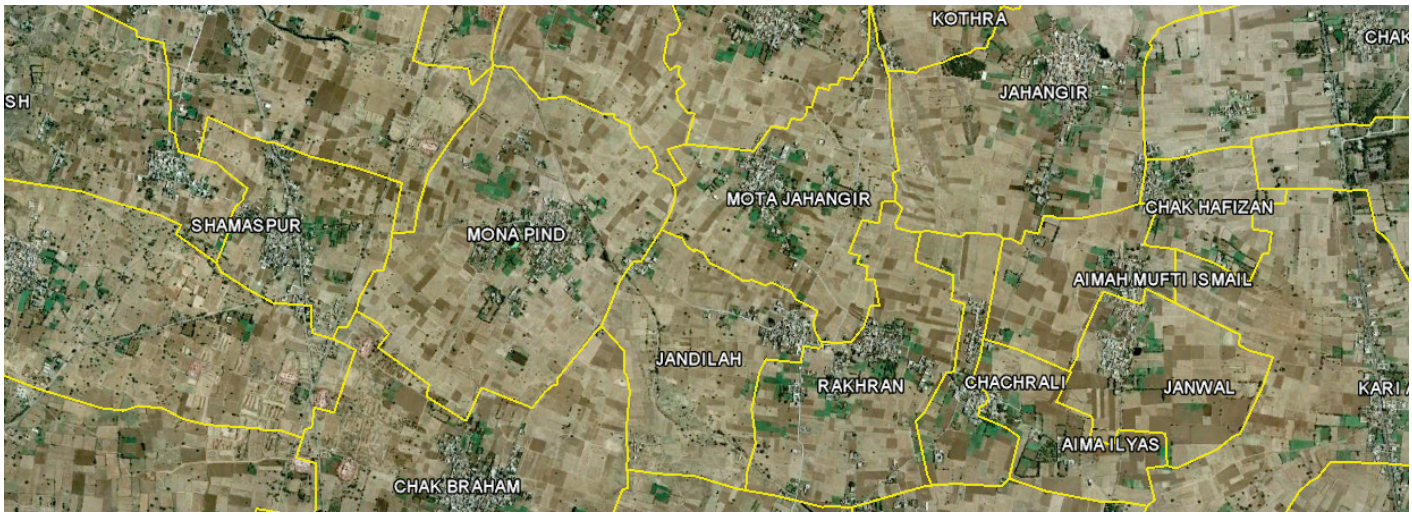



**Figure 3: The maps were stitched together like a giant jigsaw puzzle and then geo-referenced with satellite imagery by manually finding common control points. The above image shows the output: a portion of the mauza-boundary map overlaid on a satellite image.**

## 4. Preliminary Work

Mauza boundaries are defined by the Land Revenue Department. The most recent revenue maps of District Jhelum, our pilot area, were prepared in 1940. These maps, called *masavis*, are stored in the District Record Room and form the basis of all administrative boundaries at the rural level. For example, the jurisdiction of a police precinct (*thana*), electoral divisions (*halaqas*) and local government jurisdictions (*union councils*) are all defined through agglomeration of *mauza* boundaries. Yet, the maps are largely inaccessible, and some have decayed in the last 70 years to a point where they are unusable. While there could have been other approaches to extract the mauza boundaries, we identified the dire need of preservation of more than 5,000 masavis of District Jhelum as a design constraint, and started preparing a methodology for scanning these valuable documents.

Due to the age and the brittle nature of the paper on which the *masavis* are made, we did not scan them through a normal roller-scanner. Instead, we manufactured our own low-cost camera-based scanner which photographed each *masavi* in two halves. The scanning operation was completed within two weeks and generated approximately 10,000 images. We then painstakingly stitched the images together, first to combine the two halves of a *masavi* into a single image, then to mosaic the *masavis* of a whole *mauza* into one image and finally join images of adjacent *mauzas* together like a giant jigsaw puzzle. These steps required innovation in multiple directions: an understanding of the terminology of the land-revenue (*patwari*) system and their mapping practices was generated, indigenous software was developed to speed up the mosaicing process and protocols for handling missing information, torn sheets and incomplete maps were created. Finally the mosaics of *mauzas* were geo-referenced on recent satellite imagery. We observed a high degree of alignment between the 70-year old *masavis* and today's satellite images. Seeing the similarity of features such as roads, fields and streams was a source of great satisfaction for our team; it not only confirmed the high accuracy of the survey procedures employed in the 1930s, but also provided verification that the methodology that we had employed to transfer these maps to digital form was indeed correct.

While this preliminary work has demonstrated an innovative methodology to make use of historical village maps and integrating them together to yield a highly accurate administrative boundary base layer, there remain two important short-comings, which must be overcome for wide applicability of this approach:

- The method is manual and laborious, and cannot cope with large scale implementation, across an entire province. With an estimated 120,000+ *masavi* sheets in Punjab alone, the cost of completing this task will require massive human resources. The process has multiple steps, including first aligning sheets of one mauza together, then stitching multiple mauzas into a large mosaic, and finally geo-referencing it with a satellite image by marking control points. Each step requires significant time, which may become prohibitively large unless automated processing techniques are developed.

- In our previous pilot project, we only extracted the boundary of the each mauza because we were following a manual digitization process (Figure3). However, there is a wealth of information *inside* the boundaries, which includes land parcels, their identifiers and measurements, roads, land types, other features. To limit resource expenditure, we did not extract this information, which can be very useful.

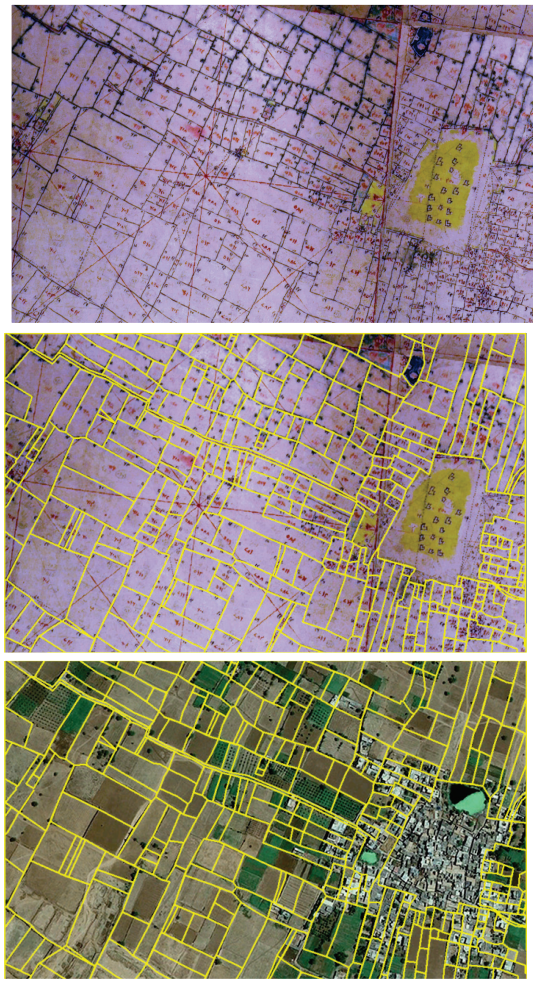


**Figure 4: [Top] A portion of a 1940 masavi. [Middle] Manually vectorized lines, shown in yellow. One of the goals of this project is to automate the vectorization process. [Bottom] Vectorized data overlayed on a recent 2010 satellite image, showing high degree of alignment.**

## 5. Proposed Work: Automatic Vectorization of Mauza-Level Maps

While we have achieved significant success in demonstrating a low-cost and scalable methodology for making a mauza boundary map like the one shown above, certain applications demand the vectorization of land-parcel data within the mauza maps themselves. By vectorization, it is meant that the lines that show land parcels and the digits that represent length are extracted symbolically, not as 'pixels' as in the current implementation, but as recognizable data (a process similar to optical character recognition or OCR). The advantage of doing so will be immense; not only will it be possible to link the land parcel data to official records but it will enable several geo-referencing applications possible, as are mentioned in the next section.

Figure 4 shows a portion of a masavi. The black lines in the top image are boundaries of *khasra* which is an individual land parcel denoting perhaps a field or a small unit of land. These form the basis of all land transactions in the country. It is our goal to detect these lines automatically in the images and mark them automatically, to generate the sort of yellow overlay shown in the middle figure. We also wish to extract the hand-written numerals on these sheets. The numerals written in black denote the length of each boundary and the ones written in red denote a unique khasra identifier, which link that particular land parcel to its cadastral record.

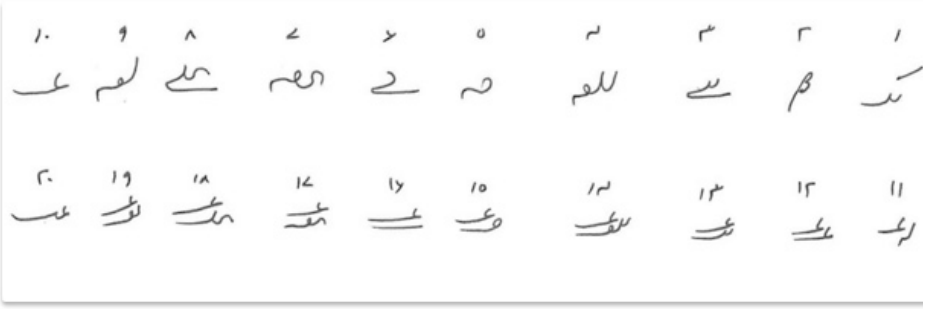


**Figure 5: Unique way of writing numerals in the Land Revenue maps, for better disambiguation. Note the variation in digits of numbers 2, 3, 4 and 6 from normal Urdu numerals. The script numerals are written on outer boundaries.**

It is estimated that there are more than 40 million *khasras* in Punjab alone. Manual marking of such a huge number of *khasra* boundaries is practically impossible. Moreover, off-the-shelf solutions are not available, because of the specialized nature of this task. This particular mapping system is unique to the Indian sub-continent, and we are not aware of any equivalent efforts yet in other countries. The way in which the numerals are written is also unique to the land-revenue system, and existing OCR systems are not trained for such scripts [See Figure 5].

**Figure 6: A zoomed portion of a masavi, showing wrinkles, torn paper, different colors, line fading and varied hand-writing, highlighting some of the scientific challenges in digitizing.**

## 6. Scientific Challenges

The images are captured with low cost, custom made camera-based scanners in unconstrained environment. Because of that, these map images contain several degradations like non-uniform illuminations, shadows, skew and curl. In addition, the sheets are often wrinkled, torn at places and have cracks and creases. These types of degradations affect human vision in correct recognition and reading of these maps. They also pose substantial challenges for vectorization of maps. The detection of lines and disambiguating them from creases and wrinkles, especially in the presence of fading and other degradations is also challenging. Recognizing hand written characters will require building an OCR system customized for this application. Finally, associating numerals and text with appropriate

lines and polygons requires an analysis of the location at which they are written, which will have to be customized based on the conventions used in this context.

## 7. Project Tasks

1) **Deshadowing and Dewarping**: Because the sheets are imaged by a low cost scanner, significant color variations are seen within and between images. Moreover, because of the old age of the sheets and the way they are stored, they are wrinkled not perfectly flat. This introduces geometric distortions in the image, which makes their perfect mosaicing difficult. Algorithms have to be developed and tested for color normalization and dewarping of geometric distortions. The IUPR group has previous research experience in both areas [see for example Bukhari, Shafait and Breuel, 2009].

2) **Line Extraction**: The sheets have two types of lines, colored red and black. Red lines form a standard 5in x 5in square grid in the background, and are very useful for aligning one sheet with another. Black lines show boundaries of individual land parcels and constitute cadastral data. A darker black line, with special 'pillar' symbols indicates the boundary of the mauza. All of these lines will have to be detected automatically.

3) **Numeral Recognition**: Like lines, there are two types of numerals on the sheets. The red numerals are writing inside each land parcel (khasra) and constitute its unique identifier. Recognizing them is necessary for linking the map data to ownership records. The black numerals actually enumerate the length of sides of a land parcel. These are to be recognized accurately to record the dimensions of a land parcel.

4) **Numeral Association**: Each numeral is associated either with a polygon or a line. Its location identifies which line or polygon it is associated with. An algorithm to associate numerals with their geometric feature will have to be developed. This is not a straight forward nearest-neighbor problem, in our observation, because of varied convention, especially in congested areas of the sheet. The orientation of numerals is also varied, and contains a cue about the line they are associated with.

5) **Data Organization**: Large amounts of data will be generated through vectorization. A proper schema will have to be designed for its storage and retrieval. Issues such as missing or unreliable recognition will have to be incorporated into this schema. Cross verification techniques, based on sequence of khasra numbers on a sheet, and collating numeric lengths with actual pixel measurements will have to be developed.

6) **Automatic Stitching and Mosaicing Application**: Based on extracted information, we will develop and test algorithms for automating the stitching and mosaicing process, which was previously done manually.

7) **Testing and Optimizing for Scalability:** We already have a dataset of 5000 sheets from District Jhelum to start the development of our algorithms. However, new sheets are also being imaged at a fast pace, and are hence, the robustness and scalability of our solution is of utmost importance. This is especially critical because mapping conventions between districts contain slight variations which will have to be addressed.

8) **Socio-Economic Mapping**: The extracted data will make a base map, which can be used to associate other datasets as attributes for the mauza, from population census, agriculture census and education data. We will develop an application to visualize these datasets on top of a geo-referenced map, as shown the figure on the cover sheet.

9) **Additional Extensions**: In addition to these core deliverables, there are numerous directions in which the project may be extended. Automatic geo-referencing based matching the extracted lines with satellite imagery may be attempted. Classification of land parcels, based on symbols and colors in the sheet will be

very useful. Finally, developing a usable application for querying, retrieving and using the available data, especially in the context of specific government department's requirement will be an important extension. These extensions are not committed within the project, but may be attempted if time allows.

## 8. Partnerships and Context

Based on the pilot project in the past year, the Punjab government had expressed high interest in the methodology. On the advice of LUMS team, the Punjab Information Technology Board, which is the primary agency dealing with e-Government in Punjab, has initiated a project to scan masavi sheets across Punjab. LUMS is already a partner in that project, ensuring access to the new raster data, and increasing the importance of coming up with a vectorization solution as quickly as possible.

This arrangement ensures that the project is not merely theoretical, but has strong practical significance and stakeholder buy-in. **A letter of support for this proposal from the Chairman of PITB is attached in the appendix.**

## 9. Examples of Benefits

While the advantages of this research for the preservation, accessibility and transparency of land revenue records are obvious, the real impact of this work lies in its applications to socio-economic mapping. Advantages of detailed mauza boundary maps have already been discussed above, but with khasra-level mapping the following additional benefits can emerge.

- *An extremely efficient strategy for geo-locating infrastructure:* The normal practice to locate, say for example, all the schools of Punjab, would be to send out teams with a GPS device and record coordinates. This is a relatively expensive strategy. If khasra locations are digitized, all one has to do is to ask the village patwari (land-revenue official) for the khasra numbers in which the schools are built.
- *A detailed ZIP-code*: One can think of khasra identifiers as a very detailed zip-code which can pin-point locations to within tens of meters, without the aid of a GPS device! Since rural citizens and officials understand the significance of a khasra, mauza and khasra identifiers could be the natural ZIP-code for Pakistan, enabling mapping of all sorts of data.
- *Time-Differential Studies*: The vectorized data will represent the state of the land in 1940, while its alignment with satellite images gives the state of the land in 2010. This 70-year differential allows one to conduct very interesting studies on urban and rural sprawl, increase in road access and change in land-use pattern to name a few. We are in a unique position to study this phenomenon, if the data can be processed efficiently.
- *Tracking of Disease and Epidemics:* The mauza is a unique spatial identifier which locates a person to a small area. Once mapped, it can help in all sorts of spatial analysis that are currently not possible. For example, the National ID card of each citizen lists their mauza (if they are from rural areas). This means that patients reporting to a hospital will be locatable on a map easily, just by their home address, allowing epidemic outbreaks, such as last year's Dengue Virus disease, to be tracked easily.
- *Land Ownership and Disputes*: Currently there is no mechanism, even for government officials, to visualize the context of a land ownership dispute. By digitization of khasra numbers, this task should become very simple, allowing powerful visualization of the context of a dispute.

- *Disaster Management:* In the recent floods, the government wanted to know the number of people affected in each area to coordinate relief efforts. While the flood footprint was easily discernable from satellite imagery, the demographics of those areas were unknown, because of the unavailability of a mauza map, which could have linked population data to the flood footprint.

These are but a few examples of the type of applications that can become a real possibility by exploiting mauza maps. Indeed, it is a pity that such a significant mapping resource has not been fully exploited in the past 70 years. By combining image processing, understanding of local context and GIS, this project aims to change the state of mapping in rural Pakistan.

## 10. Collaboration

The Image Understanding and Research Group at University of Kaiserslautern focuses on large scale, practical applications of pattern recognition, image understanding, machine learning and statistical language modeling. In particular, one of the main focus areas of the group has been in document analysis, including digitization and character recognition for large scale digital libraries using low cost digital cameras.

The Computer Vision Laboratory at LUMS School of Science and Engineering was established in 2003 as one of the first computer vision research labs in Pakistan. The group's research work is focused on both the theoretical aspects of computer vision as well as on its applications for the developing world.

Long distance collaboration facilities, including access to video conferencing equipment, Skype and dedicated bandwidth is already available at LUMS and the group has had several successful long distance collaborations in the past, leading to high quality joint publications.

Dr Khan will visit Kaiserslautern to meet face-to-face with the German collaborators for a two week duration. Prof. Breuel and Dr. Shafait are also scheduled to visit LUMS and lecture there, giving opportunity to graduate student to interact with German collaborators. Mr. Bukhari is finishing his Ph.D. in Kaiserslautern and will subsequently relocate to LUMS as a post-doctoral fellow (from April 2012), thus strengthening the collaboration significantly and acting as bridge between the two groups due to his extended presence in Pakistan.

## 11. Plan

| Project Plan (G: Germany; P: Pakistan) | Duration (Project Start: August 2011) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Year 1 | | | | Year 2 | | | | Year 3 | | | |
| | I | II | III | IV | I | II | III | IV | I | II | III | IV |
| **Preprocessing of Camera-Captured Maps** | | | | | | | | | | | | |
| De-Shadowing *(a part-time student (G))* | | | | | | | | | | | | |
| De-Warping *(a part-time student (G): 1st Quarter) + (a full-time researcher(P): 2nd Quarter)* | | | | | | | | | | | | |
| **Vectorization of Maps** | | | | | | | | | | | | |
| Line Detection *(a full-time researcher(P))* | | | | | | | | | | | | |
| Digit Extraction and Recognition *(part-time reseacher (G))* | | | | | | | | | | | | |
| Digit Association with Lines*(a full-time researcher(P))* | | | | | | | | | | | | |
| **Evaluation** | | | | | | | | | | | | |
| Ground-Truth Generation for Methods Evaluation (a part-time student (G)) | | | | | | | | | | | | |
| **Application Development** | | | | | | | | | | | | |
| Data Schemas and Organization *(P)* | | | | | | | | | | | | |
| Testing for Scalability *(G & P)* | | | | | | | | | | | | |
| Socio-Economic Mapping Application *(P)* | | | | | | | | | | | | |