A Semi-Automatic Spatial Feature Extraction Tool for Minimising Errors in GIS Data Capture*

[1]S. Mantey and [1]N. D. Tagoe
[1]University of Mines and Technology, Box 237, Tarkwa, Ghana

Mantey, S. and Tagoe, N. D., (2018), "A Semi-Automatic Spatial Feature Extraction Tool for Minimising Errors in GIS Data Capture", *Ghana Journal of Technology*, Vol. 2, No. 2, pp. 34 - 40.

Abstract

In GIS projects, data capture and maintenance are both capital and labour intensive. Over the past few years, vector data for GIS analysis have been extracted by hours of tedious and manual digitising. This method of digitising is subjective and prone to human errors. The credibility of a GIS analysis is strongly influenced by the quality of data used. The objective of this study was therefore to minimise the errors in GIS vector data capture from raster models. This was achieved by developing a semi-automatic spatial feature extraction tool. The tool is capable of extracting spatial features from orthophotos, scanned maps and high resolution satellite images within a RMSE of 0.001m. The procedures employed include: image classification, edge detection and segmentation of images as well as extraction of spatial features from raster images. Finally, the extracted features were converted to vector or GIS compatible formats such as ESRI shapefiles and other CAD formats. The result is an application capable of creating a ready-to-use vector maps by extracting spatial features from raster models within the shortest possible time. This tool is also useful for map revision and converting hard-copy maps to digital formats.

Keywords: Raster Data, Vector Data, Semi-Automatic Spatial Feature, ESRI Shapefiles

1 Introduction

GIS data is created by capturing and transforming field data into a raster or vector format, recognised by the GIS environment in which other GIS analysis is performed to address a series of environmental problems. However, since most common methods of vector data capture involve interpretation *via* the human hand, there are several types of errors that can occur. These errors undermine the reliability and efficiency of GIS analysis and projects, and may increase the cost of running GIS projects. Understanding error inherent in GIS data is critical to ensuring that any spatial analysis performed using those datasets meets a minimum threshold for accuracy. In this study, a semi-automatic spatial feature extraction tool is proposed for extracting spatial features from raster to vector. This tool would increase the cost efficiency and cost effectiveness of GIS analysis by reducing the cost and time required to extract GIS data. It also minimises the errors during the vector data capture process by enhancing the process to a semi-automatic level, thereby reducing errors caused by human limitations and interpretations. Spatial data forms the strongest component of GIS data, the effort to bring stable, accurate spatial data into GIS has always remained an important factor in any GIS. There are several ways for creating spatial data for a GIS. Each method has its strength and weakness in terms of its development and accuracies. The following provide a brief overview of some of the common methods for developing spatial data for GIS. Manual Digitising is traditionally the most common way to convert paper based spatial information (*e.g*. maps) to digital data. The paper map is mostly attached to a digitising Table. Usually between 4 or more points of which the coordinates are known are used for proper orientation and alignment. The data is then digitised by tracing the features of interest with a mouse-like handheld device called a puck. Once all the features are traced, the newly acquired data is transformed from table units to real world units using an algorithm with embedded Global Navigation Satellite Systems (GNSS) (Douglas and Peucker, 1973; Burroughs, 1986). GNSS consist of three main satellite technologies, *i.e*. GPS, Glonass and Galileo (Salgado *et al*., 2001). Each of these satellite-based technologies consists of three segments namely; space segment, control segment and user segment (Lachapelle *et al*., 2002; Feng, 2003). GNSS provides precise positioning with global coverage and allow electronic receivers to determine the longitude, latitude and elevations of respective locations (Mulassano, *et al*., 2004). The USA developed GPS consist of a constellation of more than 30 satellites. The Russian Federation developed Glonass is made up of constellation of 24 satellites. The Galileo also developed by the European Union through the European Space Agency consists of a constellation of 30 satellites. A combined constellation of GPS, Glonass and Galileo can improve satellite availability, redundancy and geometry for precise geodatasets. Geodatasets can be obtained from digital imagery. Most commonly, orthophotos and satellite imagery are utilised in a process called "Supervised classification" in which a user selects a sample of pixels for which the type of land cover is known (vegetation species, land use, *etc*.) (Dempsey, 2006). Remote Sensing software like ERDAS or ENVI uses classification algorithm to classify a

*Manuscript received December 10, 2017
Revised version accepted March 20, 2018

digital image into these named categories based on the sample pixels (Dempsey, 2006). In contrast to the other methods discussed, supervised classification results in a raster dataset. Heads-up digitising also known as on-screen digitising is becoming a popular method for digital conversion with the proliferation of low cost digital imagery and large format scanners. This method involves digitising directly on a computer of an orthorectified image such as a satellite image or an aerial photograph. The features of interest are traced from the image (Dempsey, 2006). The benefit of this over manual digitising is that, no transformation is needed to convert the data into the needed projection. In addition, the level of accuracy of the derived dataset is taken from the initial accuracy of the digital image. Heads-up digitising is also utilised in extracting data from scanned and referenced maps (Dempsey, 2006). Digitising in GIS involves conversion of geographic data from either a hardcopy or a scanned image into vector data by tracing the features. During the digitising, features from the traced maps or images are recorded in point, line, or polygon formats (Dempsey, 2006). Digitising can be in the form of manual digitising, heads-up digitising and automated digitising. Automated digitising involves using image processing software that contains pattern recognition to generate vector layers. However, since most common methods of digitising involves interpretation of geographic features via human hand, there are a number of errors that can take place in the course of capturing the data. The type of error that can take place when the feature is not captured properly is called a positional error, as opposed to attribute error where information about the feature recorded is inaccurate or false. These positional error types are outlined as follows. Dangles or Dangling nodes are lines that are not properly connected. With dangling nodes, gabs occur where the two lines should be connected. Dangling nodes also occur when a digitised polygon does not close, leaving a gap, thus creating open polygon (Dempsey, 2012).

Loops, Switchbacks and Knots are introduced when the digitiser has an unsteady hand and moves the cursor in a way that the line being digitised ends up with extra vertices and/or nodes. With loops and knots, the line folds back onto itself, creating small polygon like geometry known as weird polygon while in switchbacks, extra vertices are introduced and the line ends up with a bend (Dempsey, 2012). Undershoots and overshoots occur when the digitised line does not join or connect properly with the adjoining line. This can be resolved by setting a snapping tolerance during digitising (Dempsey, 2012; Jenks, 1981). Conversely, if the snap distance is set too high and the line endpoint snaps to the wrong node. One instance would be the presence of “cul-de-sacs” (*i.e*. dead ends) within a road GIS database (Dempsey, 2012; Jenks, 1981). Slivers are gaps in a digitised polygon layer with the adjoining polygons having gaps between them. Again, setting the proper parameters for snap tolerance is critical for ensuring that the edges of adjoining polygons snap together to eliminate those gaps. Where the two adjacent polygons overlap in error, the area where the two polygons overlap is called a sliver (Dempsey, 2012). Shape approximation is another digitising error which is seen the positional error tree. When digitising, the true shape of the feature is lost due to approximation in tracing the boundaries. Shape approximation occurs when the person digitising may not be able to zoom in to the pixel level of the image to study the true nature of the feature. Therefore, the digitising is done at a convenient scale level where all features could be seen.

2 Resources and Methods Used

2.1 Resources

The Semi-automatic Spatial Feature Extraction Tool (SSFET) was developed using the GUIDE tools in Matlab programming language. The basic controls used for the interface includes; Axes control, Edit text and label control, List boxes and Button controls. The User Interface controls were bind to series of scripts and functions that execute specific task on the controls “even callback”. The application performs a bi-level image processing operation on any input image using Canny Edge Detection Algorithm followed by “binary flagging” and filling of binary holes to obtain the features within the image in a binary (black and White) format. However, if the input image is an aerial photo, then the user will have to classify the image before processing it into a binary format. The developed user friendly interface for the SSFET is shown in Fig. 1 with controls for loading, processing and extracting spatial features. The flowchart in Fig. 2 shows the workflow of the SSFET architecture.

2.2 Methods

2.2.1 Feature Extraction

The methods used to extract features includes image data import and export, edge detection, classification, segmentation and other morphological filtering processes. A Canny Edge Detection Algorithm was used to detect the edges of features due to its large signal to noise ratio and the advantages of single edge response (Canny 1986).

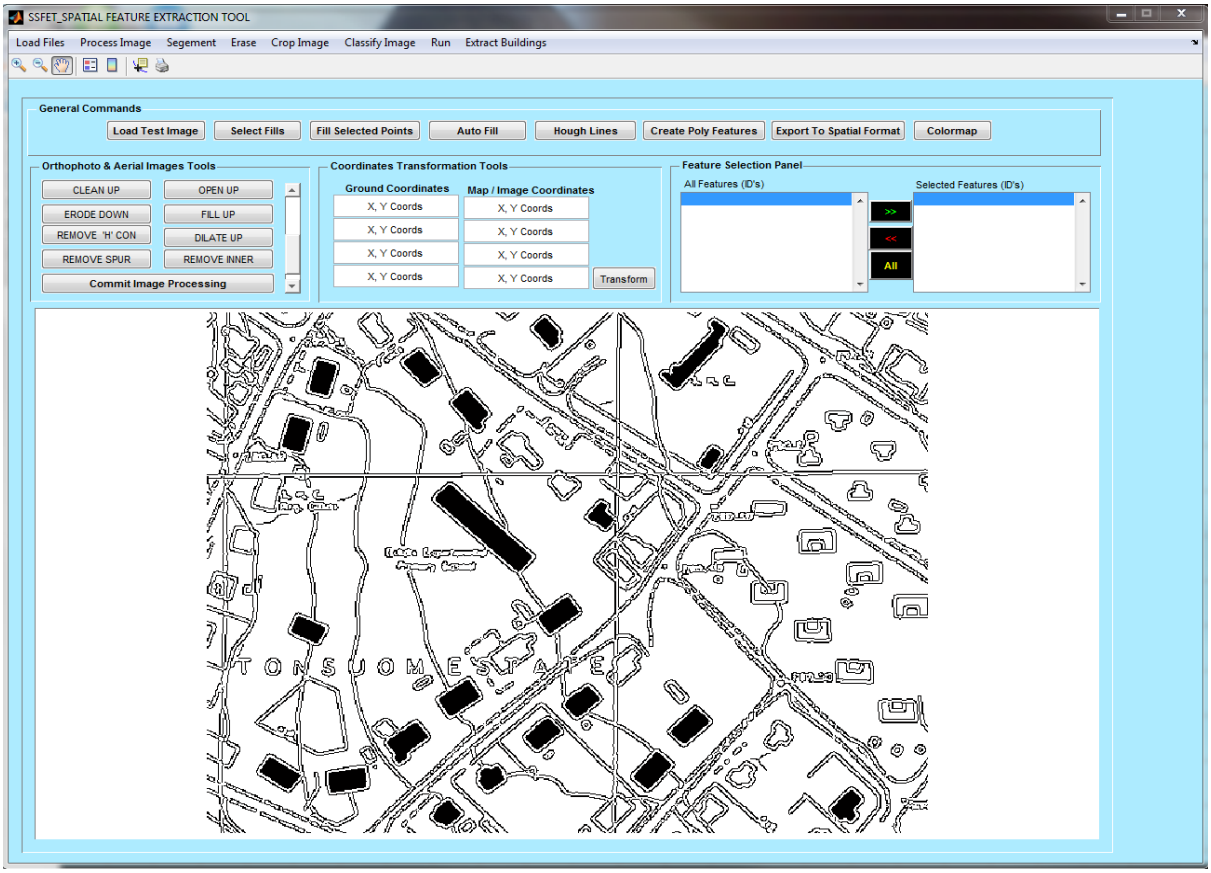

Fig. 1 SSFET's User Interface

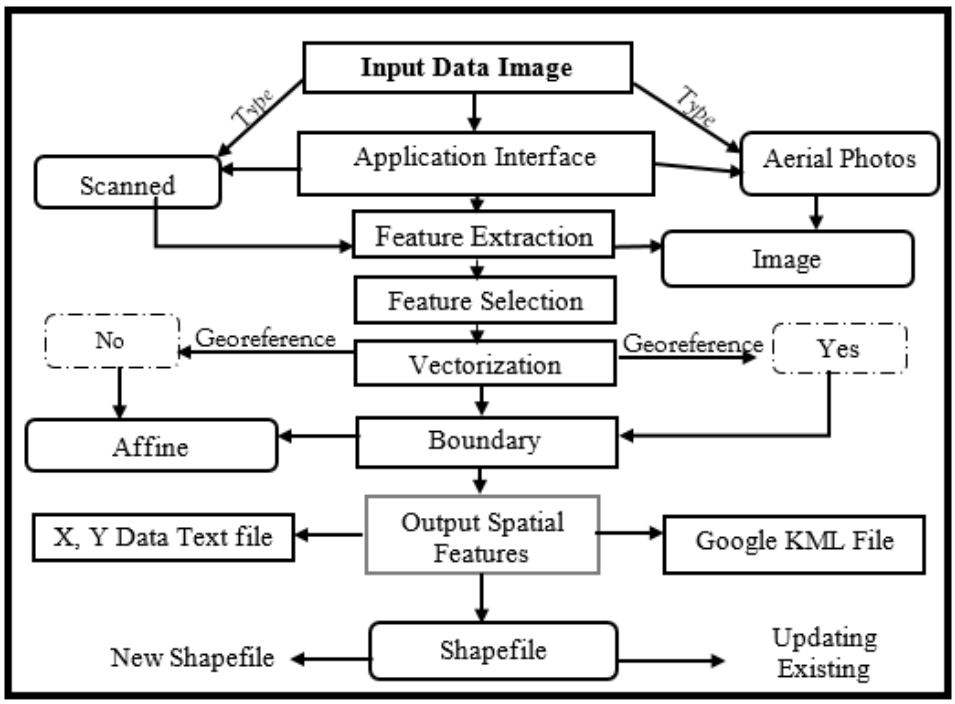

Fig. 2 Workflow of the SSFET Architecture

The processed image was then segmented into features by dividing the image area into non-overlapping and non-empty regions. The output binary image was then flagged ("binary flagging") and filled (holes filling) to obtain the segmented features (Fig. 3). If the input is an orthophoto, then image classification precedes the image segmentation process. The tool provides a classification functions which allow multiple regions to be defined for a single class, as

precaution for reducing classification errors. The classification algorithm is based on a modified Supervised Maximum Likelihood Classification (Odei, 2011).

2.2.2 Vectorisation of Features

The selected feature is processed to select centre points of pixels on its boundary that can clearly define its shape accurately without generalising the shape as seen in manual digitising. This minimises the errors incurred during manual digitising. The selected boundary points are then connected with a vector line to form the vector version of the features (Fig. 4).

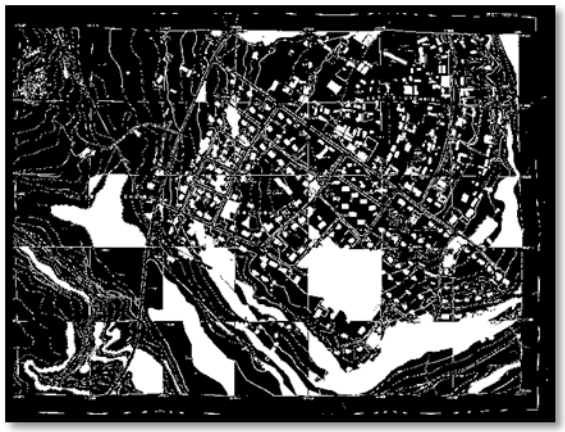

Fig. 3 Image Processed into a Bi-level Format for Vectorisation

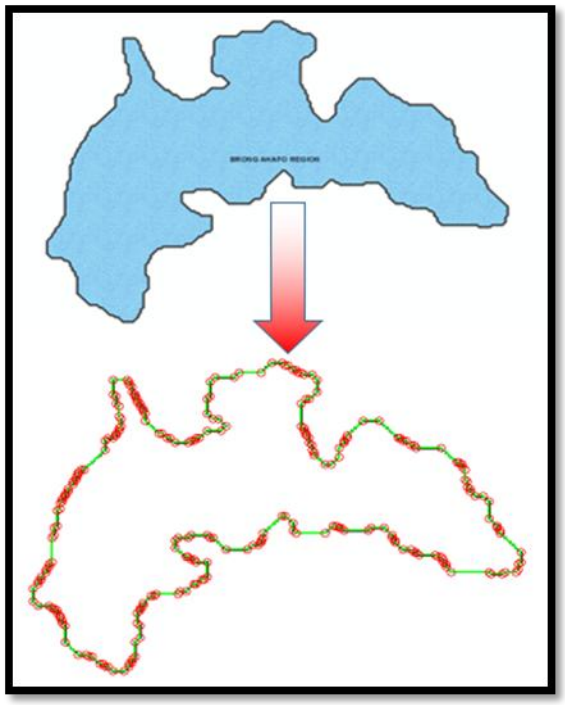

Fig. 4 Vectorisation of Selected Features

The coordinates of the selected boundary points and its unit conforms to that of the input image, therefore if the input image coordinates are the true or georeferenced coordinates, then the resulting vector will assume the same coordinates and units. However, in cases where the image coordinates are not set, the user can transform the boundary coordinates using Affine Transformation by providing a minimum of three (3) ground control points and their corresponding image coordinates to estimate the transformation coefficient which is then used to transform all other boundary points into the required coordinate system. Affine Transformation was used as it corrects for geometric distortions. This task is made easy by the application using its coordinate transformation tools which also estimate the RMS error after the transformation.

2.2.3 Spatial Feature Output

The application accepts raster data type as input and outputs a spatial data in the form of vector data which obeys topology rules of GIS (Coverage model and Geodatabase data model) (Figs. 5 and 6). The output of the application can be exported to a shapefile which is compatible with most GIS packages. It can also export the output spatial features to a KML file which is supported by most Web GIS packages and Virtual Globes like Google Earth. Moreover, the X, Y coordinate data of features can also be written down to a text file as an X, Y Data formatted as comma delimited or separated values(csv) which is supported in a wide variety of GIS and CAD packages like Autodesk AutoCAD. This application improves the accuracy of a digitised GIS data and reduces the time required for creating such data. Moreover, it provides data types that can be used in almost all GIS packages. However, the application was built to target polygon features. Though, it may be able to extract line features as well, its strength in such operations was not accessed since that was not within the scope of this study.

3 Results and Discussion

A Semi-automatic Spatial Feature Extraction Tool (SSFET) was successfully developed to extract features from scanned maps or orthorectified images in this study. In order to use the SSFET the user loads the image file (*e.g.* Scanned map) into the application using the "Load Files" button on the main menu bar. The image is processed into a binary format by inputting appropriate threshold value. Depending on the image type, the user may need to classify the image using the "Classify image" button on the main menu bar before processing the image. Fig. 5 shows an image classified with the SSFET classification tools. After processing, a black and white image (bi-level)

representation of the features on the map is displayed as shown in Fig. 7.

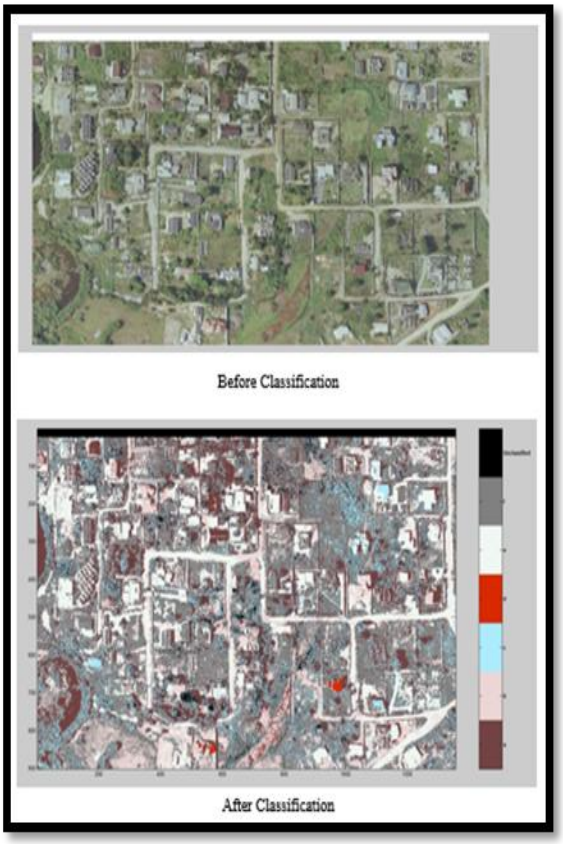

Fig. 5 Orthophoto Classified with the Application

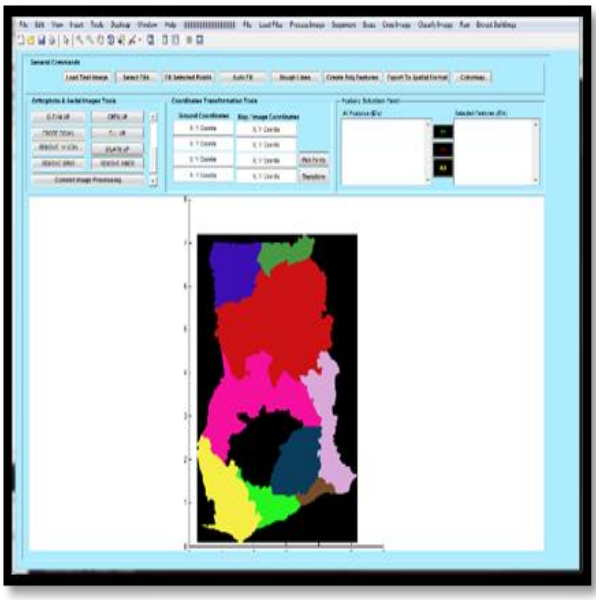

Fig. 6 The Application loaded with an Image Containing Features to be Extracted

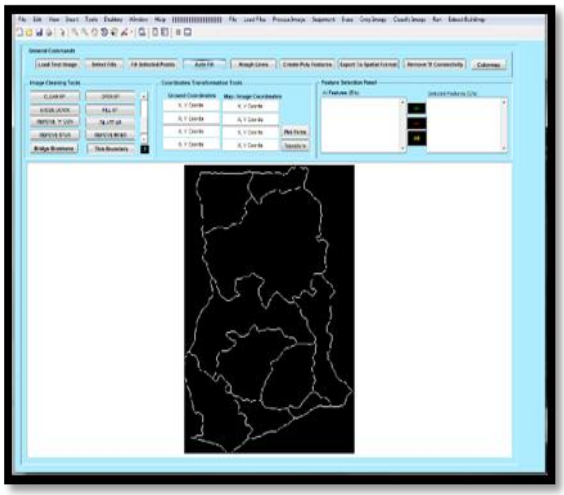

Fig. 7 Image Processed into a Bi-level Format for Feature Extraction

Features to be extracted are selected using the mouse by clicking within the features. The user may draw rectangle or freehand sketch to enclose features of interest or simply use the "Auto Fill" button on the menu bar to select all features on the map. The remaining features are "cleaned" to remove the unwanted features by clicking the "Segment" button on the menu bar. This removes all features except those selected with the mouse. To obtain the boundary of the selected features, (Fig. 8) the user clicks on "Create Poly Features", a button located on the menu bar, to create a vector boundary of all selected features. The boundary created from this tool gives a more accurate boundary free from positional errors and line generalisation.

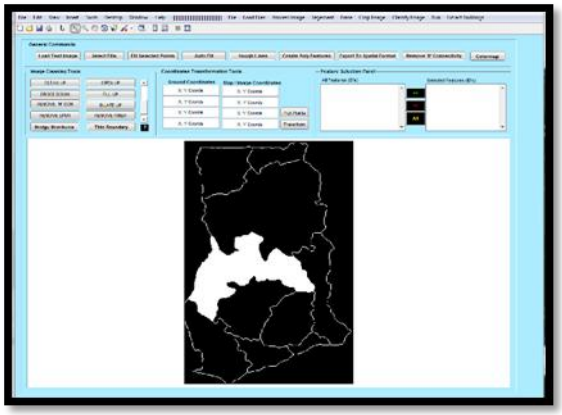

Fig. 8 Selected Feature Segmented out

After obtaining the boundary of selected features as spatial output, the user may export the features directly to GIS compatible formats like shapefile (Figs. 9, 10 and 11). The user may also choose to export an X Y Data file of the features or a KML file for sharing and visualisation on Google Earth.

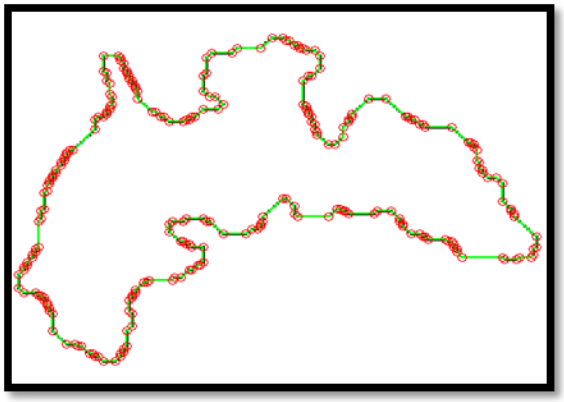

Fig. 9 Selected Feature Accurately Digitised With Minimal Errors and Time

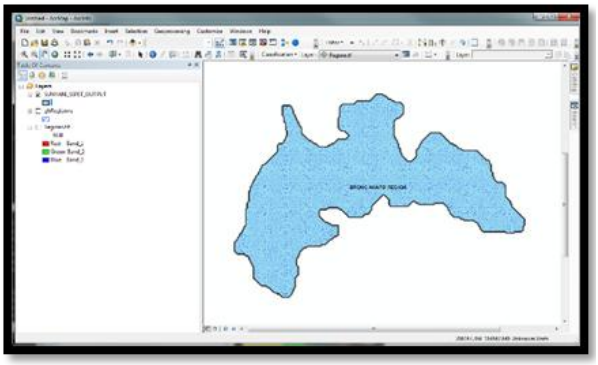

Fig. 10 Extracted Output Loaded into other GIS Platforms (ArcMap 10.0) for Validity Test

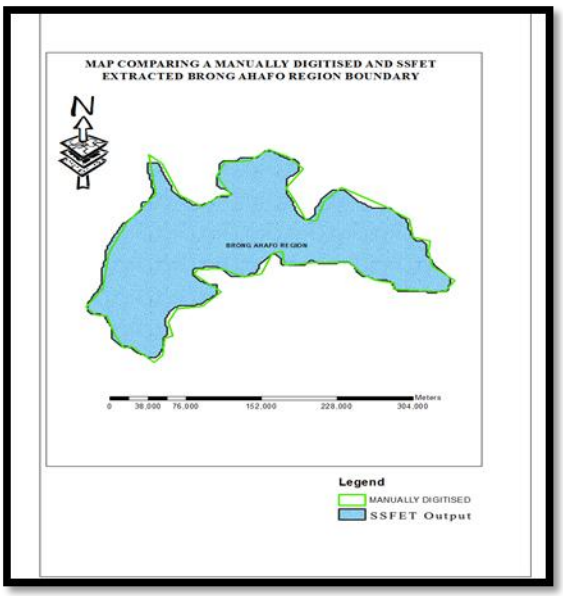

Fig. 11 Comparing output feature from manual digitising and extraction using the application

4 Conclusions

In conclusion, the semi-automatic spatial feature extraction tool developed in this study is capable of extracting spatial features from raster data models within a RMSE of 0.001m. Features extracted with this tool are more accurate relative to features extracted by manual digitising (Figs. 9, 10, 11). The extracted features can successfully be exported into a GIS compatible format such as ESRI shapefile and other CAD formats.

References

Burroughs, P.A. (1986), *Principles of Geographical Information Systems for Land Resources Assessment*, Clarendon Press, Oxford, pp.193.

Canny, J. (1986), "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 6, pp. 679-698.

David, J. B. (2013a), "GIS Data Types", *Bio Diversity GIS, www.bgis.sanbi.org/gis-primer/page_14.htm.* Accessed: March 14, 2017.

David, J. B. (2013b), "Vector Data Formats", Bio Diversity GIS, *www.bgis.sanbi.org/gis-primer/page_16.htm*. Accessed: December 14, 2017.

Dempsey, C. (2006), "Methods for Creating Spatial Databases", GIS Lounge. *www.gislounge.com/methods-for-creating-spatial-databases.* Accessed: January 20, 2018.

Dempsey, C. (2012), "Digitising Errors in GIS", *GISLounge,www.gislounge.com/digitizing-errors-in-gis*. Accessed: January 20, 2018.

Douglas, D. H. and Peucker, T. K. (1973), "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature", *Canadian Cartographer*, 10: 1 pp. 12- 122.

Feng, Y (2003), Combined Galileo and GPS: A Technical Perspective. *Journal of Global Positioning Systems*, 2 (1): pp. 67-72.

Jenks, G. F. (1981), "Lines, Computers, and Human Frailties", *American Association of Anthropological Genetics*, 7 l (1): pp. 1 – 10.

Kang-tsung, Chang, (2006), *Introduction to Geographic Information Systems*, Third Edition, McGraw-Hill, pp .506.

Lachapelle, G. M.E., Cannon, K. O'Keefe, and P. Alves (2002), How Will Galileo Improve Positioning Performance? *GPS World*, 13 (9): pp. 38 – 48.

Manuel, S. Pascual. (2011), "GIS Data: A Look at Accuracy, Precision and Types of Errors", *GIS Lounge, www.gislounge.com/methods-for-creating-spatial-databases.* Accessed: January 2, 2018.

Mulassano P., Dovis F., cllomb F. (2004), European projects for innovative GNSS-related applications. *GPS Solutions*, 7: pp. 268-270.

Oddei, G. (2011), "Image Classification Using Matlab" *Unpublished BSc Project Report,*

University of Mines and Technology, Tarkwa, pp. 16-24.
Salgado, S, Abbondanza, S, Blondel, R and Lannelongue, S (2001), "Constellation availability concepts for Galileo", *Proceedings of Institute of Navigation*, Long Beach, CA, 22-24 January 2001, pp. 778-786.
Yecheng, T. Wu. (1999), "Raster, Vector and Automated Map Digitising", *www.ablesw. com/r2v/rasvect.html.* Accessed: January 21, 2018.

Authors

S. Mantey is a Senior Lecturer at the Department of Geomatic Engineering of the University of Mines and Technology (UMaT), Tarkwa, Ghana. He holds a Bachelor of Science degree in Geomatic Engineering from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. He obtained his Master of Philosophy degree and Doctor of Philosophy from University of Cambridge and University of Mines and Technology respectively. His research interest includes application of Remote Sensing and GIS in Health and Environmental Analysis, UAVs and Web GIS applications.

N. D. Tagoe is a Lecturer at the Department of Geomatic Engineering of the University of Mines and Technology (UMaT), Tarkwa, Ghana. She was awarded BSc. degree in Geodetic Engineering at Kwame Nkrumah University of Science and Technology, Ghana. She obtained her MSc. degree in Photogrammetry and Geoinformatics from Stuttgart University of Applied Sciences, Germany. Her research interests include Close Range Photogrammetry, 3D Modelling of Cultural Heritage Sites, Remote Sensing, UAVs and Web-GIS Applications. She is a Member of IFUW.