A Note on Evaluation of Image Recognition Systems

Herbert Tesser[1] and Theron Trout[2]

[1] Marshall University, 1 John Marshall Drive, Huntington, WV, 25755, U.S.A.,
tesser@marshall.edu
[2] Strictly Business Computer Systems. Huntington, WV, 25701. U.S.A.
ttrout@sbcs.com

Abstract. The ability to systematically evaluate the results of automated image processing systems has been problematic. It remains so. In this work we address two issues in testing: 1) the use of incomplete and inaccurate reference data; and 2) we introduce a new, somewhat faster method of evaluating systems that recognize linear structures. We illustrate the latter using road and lines-of-communication recognition systems. We describe past work on evaluation systems and compare their strengths and weaknesses relative to the current work.

Introduction

This work addresses two issues germane to testing image recognition systems. The first issue is the availability of correct, comprehensive test regions. The second introduces a potentially faster way to test the correctness of linear structures.

The history of image recognition systems development is replete with examples in which testing of the systems consists of a limited number of test images created under laboratory conditions. However, systems intended for application must be tested extensively under conditions that occur in the "real-world". For example, in automated road extraction/recognition from high altitude imagery, images are captured using many sensors with diverse characteristics. Image quality, contrast, resolution, spectral content are just a few of the characteristics. A rigorous testing program must address the availability of high quality "ground truth", i.e. reference, data reflecting all of these qualities.

The development of automated methods to evaluate and diagnose image recognition systems is a significant challenge. In the absence of high quality recognition systems, authors have had to generate reference data sets by hand. These reference sets are subject to errors of interpretation.

The second concern of this work is that of performance. Traditional testing has been done for relatively small images with few roads. However, performance has not been a principal concern as the development of evaluation systems has received relatively little attention. Here we introduce a somewhat faster algorithm to perform evaluation.

Throughout this work we will illustrate our discussion with examples using automated road recognition. Roads are characterized as polylines. In many cases, it is difficult for people to identify roads from the imagery and existing databases fre-

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 60–66, 2003.

quently exhibit errors in the data that might be used for ‘ground truth”. Indeed, there are not a great number of instances where the ground truth data has been verified.

Past Work

Issue 1: Availability of correct, comprehensive test regions

Fig. 1 illustrates a common problem in the use of existing databases as references for testing purposes. Fig. 1a is an image of a suburban region and a road network overlay (blue lines). The data sets are high resolution and the vector data is derived using standard semi-automated methods. The disagreement between the imagery and the vector data illustrates the difficulty of using existing vector data sets as the basis of a testing program.

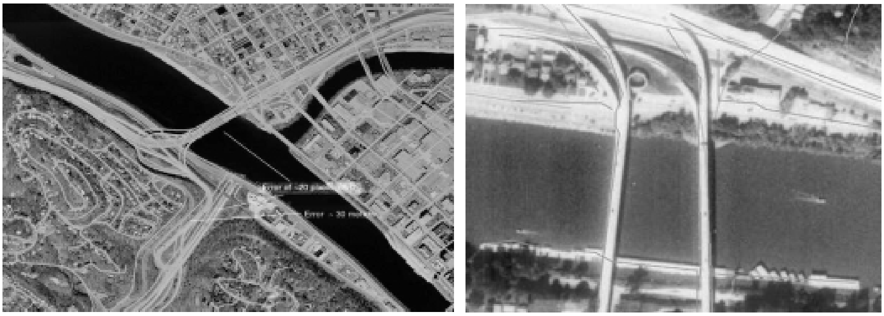

Fig. 1a. **1b.**

Fig. 1a. Image of Ohio River bridge and approaches. The image and vector data (blue lines) sets are from U.S.G.S. Error notes are added by the authors.

Fig. 1b. This example shows a case in which the reference vector data (blue lines) have completely missed one of the spans of the bridges, while the automated road recognition system (green lines) has identified both spans of the bridges. There are several other errors found in both the reference data and the automatically generated data.

To address the problem of incomplete and incorrect reference data sets, many authors have generated reference data by hand, i.e. identifying vectors using standard drawing tools. In some cases, e.g. ambiguous objects, one-lane roads in low to moderate resolution images or low contrast images, human judgment is fallible and it is problematic to use such data as reference data for evaluation.

Issue 2: A potentially faster way to test the correctness of linear structures

The evaluation system proposed and developed in [Wiederman] compares the results of the automatically generated road axes (vectors) with the reference vectors they use several principal measures including: completeness, correctness, and quality. We use completeness, false positive rate, and false negative rate. A simple transform relates the two sets of measures. They also introduce the ability to evaluate sufficiently "good" fit of the found road segments by encasing the vectors in a region or buffer. This was a significant advance in evaluation methodology.

However, the evaluation technique used in works such as [Weiderman] is O(N^2), where N is the number of vectors (road segments). Pre-sorting the polylines can reduce the coefficient of the N^2 term. However, in large data sets, dense with polylines, the time to compute the principal measures can be considerable.

Current Work

Issue 1: Availability of correct, comprehensive test regions

We contend that extensive testing of recognition systems under a broad number of environments is currently not done. Often, this is due to the lack of availability of sufficient ground truth data [Forstner]. Until such reference data is available, we propose another mode of testing, evaluation of systems using high quality - but not fully verified - reference data. Doing so implies application of statistical measures not commonly used.

We consider three distributions, ground truth (GT) – an idealized distribution, REF – the reference data distribution, and TEST – the output of the recognition system being tested.

In most work it is assumed that $P(GT^{L} \mid REF^{L}) = 1$ and $P(GT^{NL} \mid REF^{NL}) = 1$, where the superscript, L (NL), indicates the presence of a line (non-line) object. Here, we drop that assumption that the reference and ground truth are identical, and examine testing in the presence of inexact reference data.

The universe of measurement than consists of eight(8) states:

$$\begin{array}{ll} S1 = GT^{L} \; \& \; REF^{L} \; \& \; TEST^{L} & S5 = GT^{L} \; \& \; REF^{L} \; \& \; TEST^{NL} \\ S2 = GT^{NL} \; \& \; REF^{L} \; \& \; TEST^{L} & S6 = GT^{NL} \; \& \; REF^{L} \; \& \; TEST^{NL} \\ S3 = GT^{L} \; \& \; REF^{NL} \; \& \; TEST^{L} & S7 = GT^{L} \; \& \; REF^{NL} \; \& \; TEST^{NL} \\ S4 = GT^{NL} \; \& \; REF^{NL} \; \& \; TEST^{L} & S8 = GT^{NL} \; \& \; REF^{NL} \; \& \; TEST^{NL} \end{array} \qquad (1)$$

Let $P(GT^{L} \mid REF^{L})$ be the probability that a pixel is in the set of objects to be recognized under the condition that the reference data set assigns that pixel to the object set. For example, this would correspond to the case where a pixel corresponds to a road pixel when the reference data assigns the pixel to a road (non-road). Further, let $P(GT^{L} \mid TEST^{L})$ correspond to the condition that the extraction program assigns the pixel to the object set and the pixel is in the object set. This is just the measure that we are interested in calculating, along with: $P(GT^{NL} \mid TEST^{L})$, the false positive rate; $P(GT^{L} \mid TEST^{NL})$, the false negative rate. $P(GT^{NL} \mid TEST^{NL})$ is either measured or derivable from the others.

As we will see below, we measure the 4 quantities, $REF^{L(NL)}$ & $TEST^{L(NL)}$. Thus, we can calculate the joint probability distribution using Bayes' rule.

$$P(Bj \mid A) = P(A \mid Bj) P(Bj) / \sum_{i} P(A \mid Bi) P(Bi) \qquad (2)$$

In the example of road recognition, readily available vector maps with accuracy rates ~ 90% are available and can be used to extend the testing of image recognition systems to a large set of diverse image conditions.

Example

Suppose the reference data set is assumed 90% accurate, $P(GT^{L}\ |REF^{L}) = 0.9$. Then $P(GT^{NL}\ |REF^{L}) = 0.1$. Further, suppose the test program proportions are $P(REF^{L}$ |TEST~~L~~) =0.8, $P(GT^{L}\ |\ REF^{NL}) = 0.1$ and $P(REF^{NL}\ |TEST^{L}) = 0.2$. Finally, assume $P(GT^{L}) \sim 0.1$, $P(REF^{L}) \sim 0.09$, $P(TEST^{L}) \sim 0.085$.

$$\begin{aligned} P(TEST^{L}|GT^{L}) = {} & \{\ P(GT^{L}|REF^{L})\ P(REF^{L}|TEST^{L}) \\ & + P(GT^{L}|REF^{NL})P(REF^{NL}|TEST^{L})\}P(TEST^{L})/\ P(GT^{L}) \\ = {} & \{\ 0.9 * 0.8 + 0.1 * 0.2\} *0.08/0.1 \\ = {} & 0.69 \end{aligned} \qquad \mathbf{(3)}$$

Thus, if we use a reference set with 90% accuracy and the test data agrees with the reference data in 80% of the finding, we can expect that the extraction data is correct approximately 69% of the time. Conversely, assuming the reference set was identical to GT would result in claim of 90% accuracy.

The Kappa statistic is a better measure of classifier accuracy in the case where the underlying reference set is not known with absolute precision. The interclass agreement [Fitzgerald] is the ratio of the actual agreement (beyond chance) and the potential agreement (beyond chance). Here it is used to assess the evaluation system using uncertain reference data.

The Kappa statistic may be computed as follows:

Let V_{ij} be the number of pixels whose truth class is j but classified as i. The diagonal elements of the matrix V correspond to those pixels that have been correctly classified.

If $V_a = \sum_i V_{ii}$ is the overall agreement between the reference and evaluation system data and $V_c = \sum_k V_{ko}V_{ok}$ is the expected agreement by chance. Here, $V_{ok} = \sum_j V_{jk}$.

Then, the Kappa statistic is given by $K = (V_a - V_c)/(1 - V_c)$. The Kappa statistic is used to compute the combined accuracy of several classifiers [Fitzgerald], quantifying the difference between each classifier and ground truth. This is a particularly important measurement when the likelihood of chance agreement between the reference and test data sets is high. A t-test of K can then be used to evaluate the accuracy of the reference data and the test data.

Issue 2: A potentially faster way to test the correctness of linear structures

Our proposed method is image based, not vector based. We convert the vector reference data to an image, REF, by “burning in”, or scan-converting, the vector data. Similarly, the automatically extracted vectors are scan converted to an image, TEST. The images REF and TEST are binary and of the same dimensions, A = Width x Height, pixels. We generate the principal attributes – correctness, false positives rate, percent false negatives rate, using the following:

$$correctness = REF \,\&\, TEST \,/\, L_{(TEST)} \tag{4a}$$

$$completeness = REF \,\&\, TEST \,/\, L_{(REF)} \tag{4b}$$

$$false\_positives\_rate = !REF \,\&\, TEST \,/\, L_{(TEST)} \tag{4c}$$

$$false\_negatives\_rate = REF \,\&\, !TEST \,/\, L_{(REF)} \tag{4d}$$

where L(image) is the total length of the roads in image. Each of the ratios lie in the range 0.0.. 1.0.

Note that

False-positives-rate = 1 - correctness, and
False-negatives-rate = 1 - completeness

These operations in equation 4 are O(A). Usually, the area of the image is much greater than the number of vector segments. However, since the comparison among vectors is O(N^2), then it will often be the case that equation 4 will provide (slightly) better performance than those previously proposed. Improved performance can be achieved by converting the images to 1-bit/pixel format. Utilizing 32-bit integers for comparison yields in a 32-fold reduction in the number of logical operations.

Equation 1a-c puts very stringent conditions on the criteria for successful extraction of roads. We can relax those conditions by allowing tolerances associated with the reference and test data sets. Suppose we wish to count as a successful road pixel, a pixel that is within w/2 units of the reference road. Then we must scan convert the road to an image and also convert those pixels in the buffer of width w centered on the road segments. If we denote REF(w) and TEST(w) as images with the buffer and REF(1) and TEST(1) as images with no additional buffer, then the corresponding equations become:

$$correctness\_measure = REF(w) \,\&\, TEST(1) \,/\, L_{(TEST)} \tag{5a}$$

$$completenss\_measure = REF(w) \,\&\, TEST(1) \,/\, L_{(REF)} \tag{5b}$$

$$false\_pos\_measure = !REF(w) \,\&\, TEST(1) \,/\, L_{(TEST)} \tag{5c}$$

$$false\_neg\_measure = REF(1) \,\&\, !TEST(w) \,/\, L_{(REF)} \tag{5d}$$

We note that the method admits small errors in the evaluation process. In [Weiderman] the authors note “ A suitable setting of the buffer width has to consider the expected internal accuracy of the road extraction algorithm. … if the width is too large, false extractions … are considered as roads …too small .. only slightly geometrically inexact will be rejected.” They also set a maximum direction difference between the extracted vectors and the reference vectors. Here, we do not use the direction information. Hence, small errors exist when an extracted road crosses a reference road at a large angle. A second source of potential error occurs if the test road “wanders” or oscillates around the reference road.

Examples and Results

The TEST images were extracted by a fully automated road extraction system, Road-Finder Front End (RFFE). RFFE is designed to have a very low false negative rate over a wide range of image and cartographic parameters. Its principal application is as a front end of an image registration system that is sensitive to incorrectly placed road segments. In the first set of images, the reference data is drawn by hand.

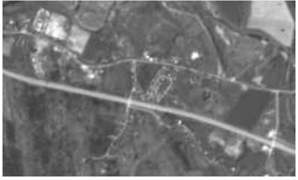

Fig. 2a. Original image

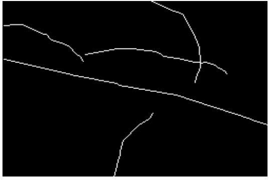

Fig. 2b. TEST extracted by the system

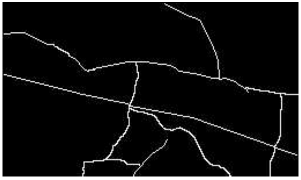

Fig. 2c. REF reference data

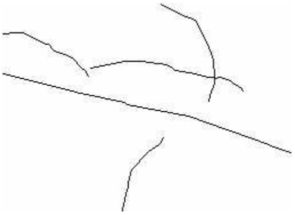

Fig. 2d. TEST image inverted

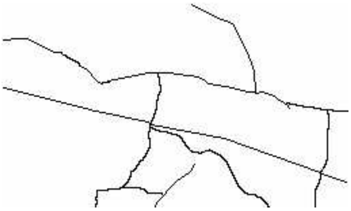

Fig 2e REF image inverted

Fig. 2a is 256x150 (=38,400) pixels of a rural area. 2b – 2e are the reference and test images and their inverted forms. The resulting data are:

Correctness = REF & TEST(1) = 61%	REF(3) & TEST(1) = 64%
False positive rate = TEST(1) & !REF(1) = 10 %	TEST(1) & !REF(3) = 3%
False negative rate = !TEST(1) & REF(1) = 30 %	!TEST(3) & REF(1) = 28%

The time to compute these quantities is less than 1 second (essentially instantaneous) on a 1.2 GHz PC running Linux. On 1024 x 866 images the computation time was ~ 2 seconds.

References

[Fitzgerald] Fitzgerald, R. W. and Lees, B. G. (1994), `Assessing the classification accuracy of multisource remote sensing data', *Remote Sensing of the Environment*, Vol. 47, pp. 362-368.

[Forstner] Forstner, W., “10 Pros and Cons Against Performance Characterization of Vision Algorihtms”, Workshop ‘Performance Characterization of Vision Algorithms’, Cambridge, 1996

[Kiiveri] Kiiveri, H. T., Caccetta, P. A., and Evans, F. H. (2001), Use of conditional probability networks for environmental monitoring. Int’l J. of Remote Sensing, Vol 22, No. 7, pp 1173-1190.

[Rosenfield] Rosenfield, G.H. (1986) Analysis of thematic map classification error matrices. *Photogram metric Engineering and Remote Sensing,* 52, 5, pp 681-686.

[Wiederman] Wiederman, C., Heipke, C., Mayer, H. Empirical Evaluation of Automatically Extracted Road Axes, CVPR Workshop on "Empirical Evaluation Methods in Computer Vision", IEEE Computer Society Press, Los Alamitos, California, 172-187

Acknowledgement

This work is supported by the Rahall Transportation Institute under the AREUS grant. We also thank the Northrop Grumman Corporation for their support.