GENETIC ALGORITHMS FOR THE UNSUPERVISED CLASSIFICATION OF SATELLITE IMAGES

Y. F. Yang [a]*, P. Lohmann [b], C. Heipke [b]

[a] Dept. of Civil Engineering, National Chung Hsing University, 250 Kuokuang Road Taichung, Taiwan 402, R.O.C - d9062503@mail.nchu.edu.tw

[b] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Nienburger Str. 1, D-30167 Hannover, Germany - (lohmann, heipke)@ipi.uni-hannover.de

Commission III

KEY WORDS: Remote Sensing Analysis, Unsupervised Classification, Genetic Algorithm, Davies-Bouldin's Index, Heuristic Algorithm, ISODATA

ABSTRACT:

Traditionally, an unsupervised classification divides all pixels within an image into a corresponding class pixel by pixel; the number of clusters usually needs to be fixed a priori by a human analyst. In general, the spectral properties of specific information classes change with the seasons, and therefore, the relation between object class and spectral cluster is not constant over time. In addition, relations for one image can in general not be extended to others. Thus, even if the number of clusters is correctly fixed for one image at one instance in time, the results cannot be transferred to other areas or epochs.

In this study, a heuristic method based on Genetic Algorithms (GA) is adopted to automatically determine the number of cluster centroids during unsupervised classification. The optimization is based on the Davies-Bouldin Index (DBI). A software programme was developed in MATLAB, - and the GA unsupervised classifier was tested on an IKONOS satellite image. The classification results were compared to conventional ISODATA results, and to ground truth information derived from a topographic map for the estimation of classification accuracy.

1. INTRODUCTION

1.1 Background on unsupervised classification

Image classification, including supervised and unsupervised classification, is an established analytical procedure of digital image processing (Lillesand and Kiefer, 2000). Supervised classification procedures require a human analyst to provide training areas, which form a group of pixels with known class label, so as to assemble groups of similar pixels into the correct classes (Avery and Berlin, 1992). In comparison, unsupervised classification proceeds with only minimal input. An unsupervised classification divides all pixels within an image into a corresponding class pixel by pixel. Typically, the only input an unsupervised classification needs is the number of classes of the scene. However, this value is usually not known a priori. Moreover, the spectral properties of specific classes within the images can change frequently and the relationships between the object classes and the spectral information are not always constant, and once defined for one image cannot necessarily be extended to others. Supervised and unsupervised classification suffers from these drawbacks.

Heuristic unsupervised classification works by establishing some mathematical model and then optimising a predefined index to determine the cluster numbers and centroids automatically. Heuristic optimization processes, therefore, are seen as a repeatable, accurate, and time-effective method to classify remote sensing imagery automatically, which is the main objective of this research. Genetic algorithms (GA) constitute one possibility for heuristic unsupervised classification. GA -have already been adopted successfully in image processing (Kawaguchi, et al., 1997), and image recognition for some special purposes such as medical treatment or criminal offence investigations (Caldwell and Johnston, 1991; Yang, et al., 2000). In this study, GA is adopted to determine number of cluster centroids of an image for use in unsupervised classification.

1.2 Status of research applying genetic algorithms

Genetic algorithms, introduced by John Holland in 1975 (Coley, 1999; Pham and Karaboga, 2000), are numerical optimisation algorithms inspired by the nature evolution process and directed random search techniques. In many fields, such as the analysis of time series, water networks, work scheduling, and facial recognition, GA have been successfully applied (Coley, 1999; Rothlauf, 2006). In 1975, De Jong (1975) executed a number of tests to study the effect of the various control parameters concerning the performance of GA. In this research, suitable values were defined, such as population size, crossover probability, and the mutation probability (Pham and Karaboga, 2000). In 2001, Bandyopadhyay and Maulik (2001) applied GA to cluster different man-made experimental point data sets and obtained very good results.

* Corresponding author. The research reported here was carried out in 2005 and 2006 while Y.F. Yang was with the Institute of Photogrammetry and GeoInformation, Leibniz University of Hannover.

2. BASES OF GENETIC ALGORITHM

The genetic algorithm is a method, which is suitable for solving an extremely wide range of problems (Coley, 1999). Recently, GA has been widely and successfully applied to optimization problems specifically in unsupervised classification of digital data sets (Ross, 1995; Bandyopadhyay and Maulik, 2002). The following sections describe the general operation of GA.

2.1 Chromosome representation

In GA applications, the unknown parameters are encoded in the form of strings, so-called chromosomes. A chromosome is encoded with binary, integer or real numbers. Since multi-spectral image data are usually represented by positive integers, in this research a chromosome is encoded with a unit (tuple) of positive integer numbers. Each unit represents a combination of brightness values, one for each band, and thus a potential cluster centroid.
The length of the chromosome, *K*, is equivalent to the number of clusters in the classification problem. *K* is selected from the range [K_{min}, K_{max}], where K_{min} is usually assigned to 2 unless special cases are considered (Bandyopadhyay and Maulik, 2002), and K_{max} describes the maximum chromosome length, which means the maximum number of possible cluster centroids. K_{max} must be selected according to experience.
Without assigning the number of clusters in advance, a variable string length is used. Invalid (non-existing) clusters are represented with negative integer "-1". The values of the chromosomes are changed in an iterative process to determine the correct number of clusters (the number of valid units in the chromosomes) and the actual cluster centroids for a given classification problem.

2.2 Chromosome initialization

A population is the set of chromosomes. The typical size of the population can range from 20 to 1000 (Coley, 1999). In the following an example is given to explain the creation of an initial population: we assume to have a satellite image with three bands, K_{min} is set to 2 and K_{max} to 8. At the beginning, for each chromosome i (i =1, 2,…,.P, where P is the size of population) all values are chosen randomly from the data space (universal data set; here: positive integers). Such a chromosome belongs to the so-called parent generation. One (arbitrary) chromosomes of the parent generation is given here:

-1 (110, 88, 246) (150, 78, 226) -1 (11, 104, 8) (50, 100, 114) -1 (227, 250, 192)

2.3 Crossover and Mutation

2.3.1 Crossover: The purpose of the crossover operation is to create two new individual chromosomes from two existing chromosomes selected randomly from the current population. Typical crossover operations are one-point crossover, two-point crossover, cycle crossover and uniform crossover. In this research, only the simplest one, the one-point crossover was adopted; the following example illustrates this operation (the point for crossover is after the 4th position):

Parent1 : ***-1 (110, 88, 246) (150, 78, 226) -1 (11, 104, 8) (50, 100, 114) -1 (227, 250, 192)***

Parent2 : (210, 188, 127) (110, 88, 246) -1 -1 (122, 98, 45) -1 (98, 174, 222) (125, 101, 233)

Child1 : ***-1 (110, 88, 246) (150, 78, 226) -1*** (122, 98, 45) -1 (98, 174, 222) (125, 101, 233)

Child2 : (210, 188, 127) (110, 88, 246) -1 -1 ***(11, 104, 8) (50, 100, 114) -1 (227, 250, 192)***

2.3.2 Mutation: During mutation, all the chromosomes in the population are checked unit by unit and according to a pre-defined probability all values of a specific unit may be randomly changed. An example explains this procedure; the bold-faced and italic units represent the result of the mutation.

Old string: (210, 188, 127) (***110, 88, 246***) -1 -1 (122, 98, 45) -1 (98, 174, 222) (125, 101, 233)

New string: (210, 188, 127) (***97, 22, 143***) -1 -1 (122, 98, 45) -1 (98, 174, 222) (125, 101, 233)

2.4 Indices identification

Based on crossover and mutation the chromosomes, once initialised, iteratively evolve from one generation to the next. In order to be able to stop this iterative process, a so-called fitness function needs to be defined to measure the fitness or adaptability of each chromosome in the population. The population then evolves over generations in the attempt to maximize the value of fitness, also called *index*.
Previous research used different indices, such as distance, separation index, Fuzzy C-Means, K-means, Davies-Bouldin Index (DBI), and Xie-Beni Index (XBI), as criteria to determine the best clustering (Ross, 1995; Bandyopadhyay and Maulik, 2002). Here, the DBI was adopted, because it is not as complex as fuzzy C-Means and one can obtain better results than with some other indices as shown using simulated data (Bandyopadhyay and Maulik, 2002; Yang and Wu, 2001). For the reasons of comparison, we also used the ISODATA algorithm.

3. METHODOLOGY

3.1 GA application of unsupervised classification

In the following paragraphs we explain the application of GA within unsupervised classification of satellite imagery. In particular, each GA procedure (such as reproduction, crossover, and mutation) is described.

3.1.1 Parent generation and population size: This procedure is an operation to produce the cluster centroids including the initial cluster centroids, which are selected randomly. This step is identical to the example given above. The range [K_{min}, K_{max}] equals [2, 8]. Two population sizes were used in or research: 40 and 100.

3.1.2 Crossover: Crossover is considered, according to the crossover probability, for example, if there are 100 chromosomes (population size 100), and the crossover probability is 0.8, the best 80 chromosomes (according to some index) are chosen for the crossover pool. The next generation (the new 100 chromosomes) are then only produced from the 80 old chromosomes of this pool.

3.1.3 Mutation: Mutation is a parameter for extending the search space; therefore, the time to reach a convergent solution increase with an increase of the mutation probability. According to the suggestion of Schaffer et al., in 1989 (Pham and Karaboga, 2000), the mutation probability is set to 0.005 here.

3.2 The Davies-Bouldin's Index

In this research, the Davies-Bouldin index (DBI) is used to represent the fitness of a chromosome (Xie and Beni, 1991; Bezdek and Pal, 1998; Swanepoel, 1999; Martini and Schöbel, 2001; Yang and Wu, 2001; Groenen and Jajuga, 2001).
First, each pixel x_n of the whole image is assigned to the nearest cluster centroid of the given chromosome, see Eq. (1):

$$\mu_{kn} = \begin{cases} 1; & \|x_n - u_k\| \le \|x_n - u_j\|, \\ 0; & otherwise \end{cases} \quad 1 \le k, j \le K; \;\; j \ne k; \;\; 1 \le n \le N \tag{1}$$

where x_n = pixel n with grey values x (one for each band)
N = total number of pixels
u_k = grey values of k^{th} cluster centroid of the previous iteration (=generation)
K = total number of clusters
μ_{kn} = membership function of each pixel x_n belonging to the k^{th} cluster

Next, the average and the standard deviation for each cluster and for the current iteration are computed (Eq. (2) and (3), followed by determining the Minkowski distance between the clusters (Eq. (4))):

$$v_k = \frac{\sum_{n=1}^{M} (\mu_{kn}) x_n}{\sum_{n=1}^{M} (\mu_{kn})} = \frac{\sum_{x_n \in X_k} x_n}{M_k} \qquad 1 \le k \le K \tag{2}$$

where v_k = average value of k^{th} cluster in the current iteration
M_k = the number of pixels belonging to the k^{th} cluster

$$S_k = \left(\frac{1}{|X_k|} \sum_{x \subset X_k} \|x - v_k\|^2\right)^{1/2} \tag{3}$$

where S_k = standard deviation of the pixels in the k^{th} cluster

$$d_{kj,t} = \|v_k - v_j\|_t \tag{4}$$

where $d_{kj,t}$ = Minkowski distance of order t between the k^{th} and j^{th} centroids. Here 2 has been chosen for t.

Subsequently, the value $R_{k,t}$ of the k^{th} cluster can be computed as Eq. (5):

$$R_{k,t} = \max_{j, j \ne k} \left\{ \frac{S_k + S_j}{d_{kj,t}} \right\} \tag{5}$$

The DB value is then defined as the average of R for all clusters in the chromosome (Eq. (6)):

$$DB = \frac{1}{K} \sum_{k=1}^{K} R_{k,t} \tag{6}$$

$$Min \quad DBI = 1/DB \tag{7}$$

The goal for achieving a proper clustering is to minimize the DBI (Eq. (7)). Thus, the fitness function for chromosome j is defined as $1/DB_j$, which is equivalent to the clustering with the smallest inner-cluster scatter and the largest cluster separation.
After calculating the DBI of each chromosome of a given population, the best chromosome is compared to the best one of the previous generation (iteration). The termination condition for the iterations is that the difference between these two values lies below a pre-defined threshold. If this condition is not met, the best chromosomes are selected into the crossover pool (see above) and a new iteration is started. The computations are also stopped once a maximum number of generations is reached.

3.3 Influence of crossover and mutation probabilities

There are five factors that influence the result of a GA algorithm: the encoding form (binary, real number and so on), the size of the initial population, the fitness function, the genetic operations (such as the one-point crossover, two-points crossover, etc.), and the probabilities for crossover and mutation (Pham and Karaboga, 2000). In this research, variations of the initial population size and the crossover probability are discussed.

3.4 Image data, ground truth and error matrices

For our research we used a multi-spectral IKONOS image. The image depicts Chandlers Ford in the U.K. and, was taken on 2000/08/25 with 4 meters pixel size and 11 bits per pixel (see Figure 1). We used a subset with a total of 18330 pixels. A higher resolution map served as a reference for obtaining ground truth information.
We measure classification success using the well-known criteria *producer's accuracy or completeness* (the number of pixels that are correctly assigned to a certain class divided by the total number of pixels of that class in the reference data) and *user's accuracy or correctness* (the number of pixels correctly assigned to a certain class divided by the total numbers of pixels automatically assigned to that class).

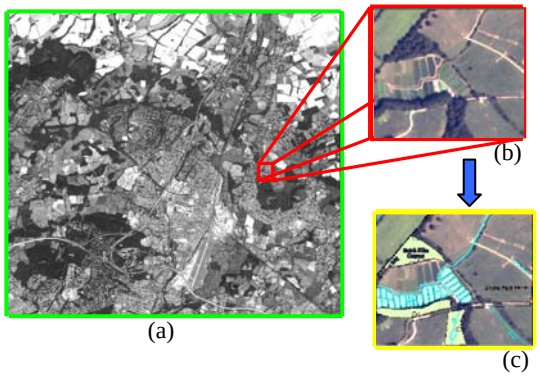


Figure 1. (a) The original IKONOS image; (b) The extracted IKONOS subset image; (c) Ground truth map superimposed on the subset image.

4. ANALYSIS RESULTS

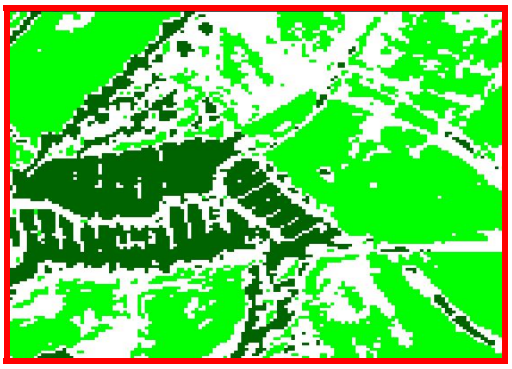
In this section, we present the results of our research. The following parameters of the GA classifier were set:

1.	chromosome length	8
2.	single point crossover	
3.	crossover probability	0.4 and 0.8
4.	population size	40 and 100
5.	mutation probability	0.005

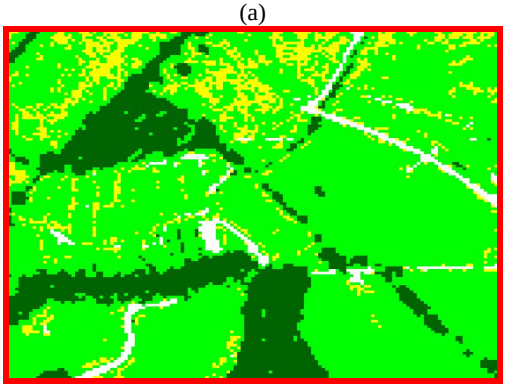
In the ground truth data four distinct classes can be found: *road*, *farmland*, *forest*, and *others*. Figures 2 and 3 show the results with one colour per class: *road* in white, *farmland* in light green, *forest* in dark green, and *others* in yellow. The error matrices of the four experiments are shown in Tables 1 to 4.

Compare Figure 2 (a) with Figure 2 (b) and Table 1 (a) and (b), when the population size increases, the overall accuracy increases from 49.1% to 69.8% and four instead of only three classes are found. The same effects are evident from Figure 3 (a) and 3 (b) and Table 2: the overall accuracy increases from 54.4% to 71.1% and again four classes can be detected with a population size of 100. When comparing the effect of the two investigated parameters, it is clear that the population size is significantly more important than the mutation probability. With a few exceptions, most notably the completeness of roads, the producer's and the user's accuracy all increase when increasing the population size.

As a reference, Figure 4 and Table 3 depict the results of the traditional ISODATA with four classes as prior information. The results of the GA are better (taking the higher population size) than the ISOADATA results; it should be mentioned, however, that the computational expense for GA is significantly larger than that for the ISODATA algorithm.



(a)



(b)

Figure 2. Results with (a) population size 40, and crossover probability 0.4; (b) population size 100 and crossover probability 0.4

		Reference Data			
		Road	Farmland	Forest	Other
Classification	Road	77.7%	31%	49.7%	52%
	Farmland	7%	52.6%	12.6%	48%
	Forest	15.3%	16.4%	37.7%	0%
	Other	0	0%	0%	0%

Producer's Accuracy (Completeness)	**User's Accuracy (Correctness)**
Road=77.7%	Road=14%
Farmland=52.6%	Farmland=87.4%
Forest=37.7%	Forest=33.5%
Other=0%	Other=0%
Overall accuracy=49.1%	

(a)

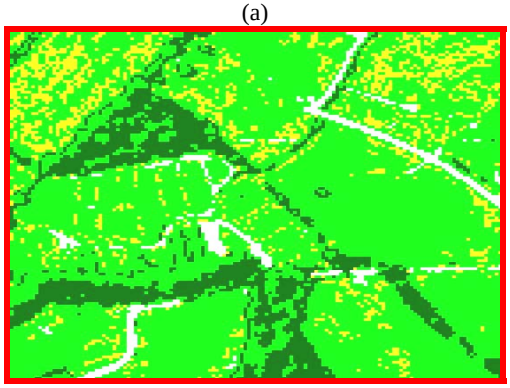
		Reference Data			
		Road	Farmland	Forest	Other
Classification	Road	33.3%	5.5%	1.4%	38.6%
	Farmland	50.2%	76.5%	22.9%	54.7%
	Forest	5.3%	6.9%	74.7%	0.4%
	Other	11.2%	11.1%	1%	6.3%

Producer's Accuracy (Completeness)	**User's Accuracy (Correctness)**
Road=33.3%	Road=35.9%
Farmland=76.5%	Farmland=81.4%
Forest=74.7%	Forest=79.3%
Other=6.3%	Other=0.8%
Overall accuracy= 69.8%	

(b)

Table 1. (a) and (b). Error matrices for results depicted in Figure 2

(a)



(b)

Figure 3. Results with (a) population size 40, and crossover probability 0.8; (b) population size 100 and crossover probability 0.8

		Reference Data			
		Road	Farmland	Forest	Other
Classification	Road	77.1%	33.2%	5%	88.8%
	Farmland	19.2%	57.6%	42.3%	7.2%
	Forest	3.7%	9.2%	52.6%	0%
	Other	0%	0%	0%	0%

Producer's Accuracy (Completeness)	**User's Accuracy (Correctness)**
Road=77.1%	Road=16.4%
Farmland=57.6%	Farmland=82.9%
Forest=52.6%	Forest=56.7%
Other=0%	Other=0%

Overall accuracy=54.4%

(a)

		Reference Data			
		Road	Farmland	Forest	Other
Classification	Road	38.4%	3.2%	1.9%	45.3%
	Farmland	49.6%	79.5%	39.2%	48.9%
	Forest	7.3%	4.5%	57.9%	0%
	Other	4.7%	12.7	1%	5.8%

Producer's Accuracy (Completeness)	**User's Accuracy (Correctness)**
Road=38.4%	Road=44.4%
Farmland=79.5%	Farmland=84.5%
Forest=57.9%	Forest=72.8%
Other=5.8%	Other=0.6%

Overall accuracy= 71.1%

(b)

Table 2. (a) and (b). Error matrices for results depicted in Figure 3

Figure 3. Results of ISODATA algorithm (4 clusters)

		Reference Data			
		Road	Farmland	Forest	Other
Classification	Road	64%	13.9%	52.2%	69.1%
	Farmland	27.2%	70%	26.4%	9%
	Forest	8.4%	12.5%	20.7%	21.5%
	Other	0.4%	3.6%	0.7%	0.4%

Producer's Accuracy (Completeness)	**User's Accuracy (Correctness)**
Road=64%	Road=17.5%
Farmland=70%	Farmland=88.7%
Forest=20.7%	Forest=33.6%
Other=0.4%	Other=0.2%

Overall accuracy= 65.1%

Table 3. Error matrix from ISODATA results

5. CONCLUSION

One of the a priori inputs traditionally needed for unsupervised classification is the number of clusters in the data set. In many cases, however, this number of classes is not available. This research describes a procedure for unsupervised classification based on genetic algorithms, which is able to estimate the required number of clusters as part of the procedure. In order to evaluate the individual results we used the Davues-Bouldin's Index (DBI).

The effectiveness of the new technique was evaluated using examples of IKONOS satellite image data. Based on independent ground truth an overall accuracy of 71.1% was reached as compared to 65.1% when using the ISODATA algorithm. For a number of applications this accuracy is still acceptable.

GA has a number of free parameters. Two of them, namely population size and the crossover probability were considered in this research. In our results the population size proofed to be significantly more important than the crossover probability. In future research we will further investigate the potential influence of the other parameters and also consolidate our results using more test data and alternative indices for measuring the chromosome fitness.

6. REFERENCES

Avery, T.E. and Berlin, G.L., 1992. *Fundamentals of remote sensing and airphoto interpretation.* MacMillan Publishing Company, New York, 472 p.

Bandyopadhyay, S., and Maulik, U., 2001. Nonparametric genetic clustering: comparison of validity index. *IEEE Transactions on systems man, and cybernetics-part C: Applications and reviews*, 31(1), pp.120-125.

Bandyopadhyay, S., and Maulik, U., 2002. Genetic clustering for automatic evolution of clusters and application to image classification. *IEEE pattern recognition*, Vol.35, pp.1197-1208.

Bezdak, J.C., and Pal, N.R., 1998. Some new indexes of cluster validity. *IEEE Transactions on systems, man, and cybernetics*, part B, 28(3), pp.301-315.

Caldwell, C., and Johnston, V.S., 1991. Tracking a Criminal Suspect Through "Face-Space" with a Genetic Algorithm. *Proceedings of the 4th International Conference on Genetic Algorithms*, Morgan Kaufmann, pp.416-412.

Coley, A D., 1999. *An Introduction to Genetic Algorithms for Scientists and Engineers.* World Scientific, Singapore, 188p.

De Jong, K.A., 1975. *An analysis of the behavior of a class of genetic adaptive systems*. Ph.D dissertation, University of Michigan, Ann Arbor, Michigan.

Groenen, P.J.F., Jajuga, K., 2001. Fuzzy clustering with squared Minkowski distances. *Fuzzy Sets and Systems*, Vol.120, pp.227-237.

Kawaguchi, T., Baba, T., Nagata, R., 1997. 3-D object recognition using a genetic algorithm-based search scheme, *IEICE transactions on information and systems*, E80D(11), pp.1064-1073.

Lillesand, T.M. and Kiefer, R.W., 2000. *Remote Sensing and Image Interpretation.* John Wiley & Sons, New York. 724 p.

Martini, H., and Schöbel, A., 2001. Median and center hyperplanes in Minkowski spaces -- a unified approach. *Discrete Mathematics*, Vol.241(1), pp.407-426.

Pham, D.T., and Karaboga, D., 2000. *Intelligent Optimisation Techniques*. Springer, London, Great Britain, 261p.

Ross, T.J., 1995. *Fuzzy logic with engineering applications.* Mc Graw-hill, Singapore, 592p.

Rothlauf, F., 2006. *Representations for Genetic and Evolutionary Algorithms*. Springer, Netherlands, 314p.

Swanepoel, K.J., 1999. Cardinalities of k-distance sets in Minkowski spaces. *Discrete Mathematics*, 197(198), pp.759-767.

Xie, X.L., and Beni, G., 1991. A Validity Measure for Fuzzy Clustering. *IEEE Transaction on Pattern Analysis and Machine Inteligence*, 13(8), pp. 841-847.

Yang, G., Reinstein, L.E., Pai, S., Xu, Z., Carroll, and D.L., 2000. A new genetic algorithm technique in optimization of prostate implants. *Medical Physics*, 35(5), pp.104-112.

Yang, M.S., and Wu, K.L., 2001. A new validity index for fuzzy clustering. *IEEE International Fuzzy Systems Conference*, pp.89-92.

7. ACKNOWLEDGEMENTS

We would like to thank the Ordnance Survey of United Kingdom to offer the IKONOS image data within the OEEPE test and National Scientific council (NSC) of Taiwan Taipei for subsidizing this research.