Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 21 - 23, 1977

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
77CH1218-7 MPRSD

SEARCH - AN EFFICIENT, AUTOMATIC TRAINING SAMPLE SELECTION ALGORITHM

RONNIE W. PEARSON
Lyndon B. Johnson Space Center
Earth Resources Laboratory
1010 Gause Boulevard
Slidell, Louisiana 70458

Program SEARCH is an unsupervised trainer for any maximum likelihood classifier. The need for such a program was recognized because of the cost involved in acquiring sufficient, reliable training information over very large areas. The use of existing known clustering algorithms was ruled out for two reasons: 1) insufficient speed, and 2) most importantly, the inability to derive and use the off-diagonal elements of the covariance matrix in data with known high correlation of successive channels.

The general approach selected to satisfy the need and avoid the two listed hazards is as follows: 1) divide the survey area into six scan by six element areas, 2) analyze each area as a possible training sample, 3) store signatures for all areas that "seem to be homogeneous", 4) once fifty signatures have been stored, merge the pair having the smallest pairwise divergence, thus reducing the number of signatures by one, (repeat steps 2, 3, and 4 through the data set), and finally 5) merge resultant signatures using divergence specified by the user at run time.

What does "seem to be homogeneous" mean? It seems obvious that the desired criteria would be a multivariate normal distribution test. After inspecting several thousand sets of training sample statistics, means, covariance matrices and histograms, and not finding adequate data to support this test the author chose to place lower and upper bounds on the standard deviation of each channel. The lower bound precludes extremely peaked signatures that tend to generate extremely high divergencies with most any other signature. The upper bound is chosen to insure homogeneity within the area. The upper bound required to give homogeneity in the lower radiometric values of LANDSAT must be increased to obtain any data at the higher radiometric values, probably due to the type decompression applied to LANDSAT data. In processing ten LANDSAT frames, typical bounds that have proven satisfactory in the Southeastern United States are .7 for the lower bound, and the greater of 1.2 and 6 percent of the mean for the upper bound. Different bound selections may be required in other areas. The average cost for running SEARCH on a LANDSAT frame is $150 (2 hours) on a Varian mini-computer. The resulting classified images appear very detailed with good definition. Actual analysis of random points for cluster naming and accuracy verification is now in progress.