MULTI-SENSOR DATA FUSION TECHNIQUE: PART II

Sharon Rodier[(1)], Yongxiang Hu[(2)], Mark Vaughan[(3)], Dennis Hlavka[(4)] , Tom Arnold[(5)]

[(1)] SAIC, MS 927, Langley Research Center, Hampton VA, USA 23681; E-mail: s.d.rodier@larc.nasa.gov
[(2)] NASA, MS 475, Langley Research Center, Hampton VA, USA 23681; E-mail: Yongxiang.Hu-1@nasa.gov
[(3)] SAIC, MS 475, Langley Research Center, Hampton VA, USA 23681; E-mail: m.a.vaughan@larc.nasa.gov
[(4)] SSAI, Goddard Space Flight Center, Code 613.1 Greenbelt, MD USA 20771; E-mail: sgdlh@virl.gsfc.nasa.gov
[(5)] SSAI, Goddard Space Flight Center, Code 613.2, Greenbelt, MD USA 20771; E-mail: arnold@climate.gsfc.nasa.gov

ABSTRACT

Recent clustering studies with the Kohonen self-organizing map (SOM; [1]) have shown that combining collocated measurements from the MODIS Airborne Simulator (MAS; [2]) and the Cloud Physics Lidar (CPL; [3]) can produce a greater number of distinct classes of data (i.e., scene types) than can be derived from either measurement alone. This is because the addition of the lidar parameters to the SOM feature vector enables more accurate classification and labeling of the derived clusters [4,5]. In this paper, we describe a technique for extending the classifications derived from the nadir track analysis, where we have coincident measurements from both instruments, to the full passive sensor swath, for which we have only MAS measurements. Preliminary validation studies show that we can expect a classification success rate of better than 70% when applying this method.

1. INTRODUCTION

Thin clouds and water vapor in the upper troposphere absorb thermal radiation from the surface and re-emit thermal radiation to space at a significantly lower temperature, thus exerting a strong greenhouse effect on the climate. When making measurements over land surfaces, passive sensors such as AIRS and MODIS have difficulty distinguishing upper troposphere water vapor from thin clouds, since water vapor and thin clouds are similar in their absorption and emission of thermal radiation, and the solar radiation signature of the thin clouds is weak when compared to the surface reflectance. This inability to identify the thin clouds over land surfaces has deleterious affects on the AIRS temperature and moisture profile retrievals. Space-based lidar excels in the detection of thin cirrus. However, its nadir-only viewing geometry limits the extent to which these observations can be used to immediately improve the AIRS and MODIS retrievals.

The goal of the research described here is to use data fusion techniques to extend the knowledge gained from coincident lidar and passive sensor measurements to the entire passive sensor swath. Our tool of choice is the SOM, an artificial intelligence technique that provides a robust method for identifying clusters of similar data items. When implemented on the appropriate hardware, SOM algorithms can analyze large quantities of data in a relatively short time. In preparation for the deluge of data that will be delivered following the launch of CALIPSO [6], custom-tailored software has been developed to adapt the SOM algorithm for use in data fusion tasks that ingest measurements from multiple remote sensing platforms. The current version of this software employs CPL and MAS measurements as proxies for space-based instruments (CALIPSO and MODIS, respectively).

The specifics of using the SOM as a clustering tool have been described in detail elsewhere [2,4,5], and will not be repeated here. Instead we focus on the progress we have made in (a) expanding the amount of data used to determine the cluster centers; (b) validating the classification of nadir track data; and (c) using the multi-instrument cluster characteristics as the basis for a classification scheme that is applied to the MAS-only swath data that was not used in the cluster analysis.

In the discussion that follows, it is important to draw a clear distinction between clustering and classification. The activity of clustering involves partitioning a data set into a number of discrete groups (clusters) according to some measure of similarity. Classification, on the other hand, attempts to categorize here-to-fore unanalyzed data points based on their similarity to a predetermined set of clusters.

2. DATA VOLUME AND PREPARATION

Put simply, using more data on the input side of the SOM means better, tighter, and more physically meaningful clusters on the output side. To prepare for high volume, space-based data streams, we have modified our software to substantially increase the amount of data that can be used in any single analysis. The data we used to test our updated system was acquired as part of the Terra-Aqua Experiment, which was conducted in November and December of 2002 (TX-2002; see http:// cimss.ssec.wisc.edu/tx2002/). For this campaign, both MAS and CPL were mounted on NASA's ER-2, and the aircraft conducted numerous surveys over the Gulf of Mexico and the southern Great Plains of the United States. From the wealth of data collected, we selected all of the daytime measurements acquired over land and inland water bodies (e.g., lakes, ponds, and streams). The CPL

portion of the data used in this analysis is illustrated in Fig. 1. As indicated on the plot, this data was acquired on November 24, 27, and 29, and December 7, 11, and 12 of 2002. Together with the collocated MAS data [5], 49800 CPL profiles were used. This total is equivalent to approximately 14 hours of data.

For the clustering analyses, the SOM feature vector contained 17 parameters. Nine of these were derived from CPL measurements, and the remaining eight were from MAS. Based on the ISCCP Cloud Height Classification Scheme [4], each CPL profile was divided into three altitude regimes. Using layer boundaries recorded in the CPL data files, low-, middle-, and high-altitude values for the 532 nm layer-integrated attenuated backscatter, and the 1064 nm parallel and perpendicular layer-integrated attenuated backscatters were calculated from the lidar backscatter data. The eight parameters derived from MAS data were the reflectances at 0.55 um, 0.65 um, and 8.7 um; the ratio of reflectances between 1.6 um and 0.55 um (1.6um/0.55um), and between 2.2 um and 0.55 um (2.2um/0.55um); the brightness temperatures at 3.7 um and 12 um; and the difference between the brightness temperatures at 11 micron and 12 micron. (Note that, in an effort to reduce misclassifications, the feature vectors constructed for this analysis used more and somewhat different MAS parameters than the vectors used in the analyses described in [4].)

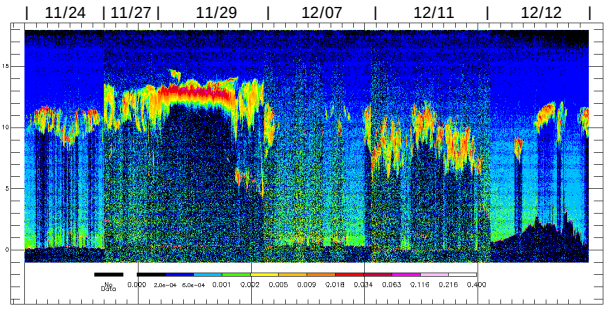

Fig. 1: CPL attenuated backscatter coefficients at 532 nm acquired during 6 days of the TX-2002 campaign.

3. VALIDATING THE MAS-ONLY NADIR-TRACK CLASSIFICATIONS

The approach we use to extend the data rich nadir-track measurements to the full passive sensor swath is actually quite straightforward. Using only the eight MAS parameters from the 17-element cluster centers determined by a SOM analysis of the nadir track data, we can apply a distance metric to determine the "closest match" between the MAS swath pixels and the classes represented by the SOM cluster centers. To be useful, however, the accuracy of all aspects of this technique must be well validated. To this end, we designed a set of test cases that allow us to evaluate and validate the correctness of our classification scheme. These test cases were constructed by separating the reference data set (i.e., the 49800 collocated nadir track measurements) into matched training sets (containing N vectors) and test sets (containing 49800-N vectors). For each test, the training set data was used as input to a SOM analysis that yielded eight cluster centers. Each cluster center is characterized by a 17-element feature vector, as described in Section 2. The test set data was then classified using a normalized Euclidean distance metric. Two types of classification were performed. The first used the full 17-element feature vector, while the second used only the MAS components of the vector (that is, only 8 elements instead of the full 17).

Three tests were conducted. For the first test, we randomly selected and removed 5% of the original 49800 data points. The removed data serve as an independent test set used to validate the classification scheme. The remaining data, known as the training set, served as input to the SOM clustering algorithm. For the second test, an arbitrarily selected subset of five thousand contiguous records was removed and set aside for testing. The third test was identical to the second, except that in this trial ten thousand contiguous records were removed and set aside. For each test, the SOM analysis was conducted using the same procedure. We first specified a hexagonal shaped, 4 x 2 map, which produce eight clusters. The SOM analysis then takes place in two phases: the ordering phase and the converging phase [1]. The ordering phase consisted of 2,500,000 cycles, with the learning rate initialized to 0.08, and the initial radius set to 10. The converging phase consisted of 25,000,000 cycles, with a learning rate of 0.04 and a final radius of 2. Finally, we also used the same procedure to conduct an additional reference test, in which all 49800 vectors were used in the SOM analysis. For this test there was no associated set of test data.

The first critical step in validating our classification scheme is to examine the test sets processed using the 17-element cluster centers, and determine the degree to which the feature vectors were properly allocated to the correct class. For supervised learning algorithms such as a back-propagation neural network, this analysis is a simple exercise, because the correct classification of the test set data is known a priori. Supervised learning algorithms are trained to recognize known patterns in a data set. The SOM, however, is an example of unsupervised learning, and these routines were devised specifically to discover here-to-fore unrecognized patterns. As a consequence, determining whether a test vector has been properly allocated to the correct class is somewhat tricky, because the "correct classes" are only known after the fact, and we have no a priori knowledge on which to base a comparison. Therefore, to assess the performance of the classification scheme we

have assumed that the cluster centers determined by the reference test represent "truth". Since the reference test partitions all of the data into one of eight clusters, we can decide whether a feature vector from the test data has been properly classified by (a) ascertaining its class membership as determined by the reference test, and then (b) comparing that class membership to the class assignment derived from the training set data. In all cases, the class characteristics are sufficiently distinct so that no confusion arises in determining class identity between different test runs. Based on this procedure, we achieved an exact classification match for 98% of the feature vectors in test cases 1 and 2. However, due to data decimation of the training sets, results for case were substantially lower (69%). Tables 1 through 3 report the distribution of feature vectors allocated to the eight clusters for each of the SOM analyses, and present the classification results that were derived by each of the two classification schemes.

Table 1: Cluster assignments for test #1 (5% random removal); for the 17-element classifications (C-17), columns are shown for the training set (training) and for the test set (test); the 8-element classification (C-8) was applied to all 49800 feature vectors. Numbers in parentheses are normalized values describing the fraction of the whole represented by each cluster.

Cluster	C-17 training	C-17 test	C-8, all data
1	13231 (0.28)	687 (0.28)	14531 (0.29)
2	972 (0.02)	46 (0.02)	1414 (0.03)
3	5086 (0.11)	273 (0.11)	1685 (0.03)
4	4334 (0.09)	235 (0.09)	8485 (0.17)
5	5724 (0.12)	272 (0.11)	6414 (0.13)
6	5379 (0.11)	273 (0.11)	7765 (0.16)
7	3211 (0.07)	201 (0.08)	2281 (0.05)
8	9383 (0.20)	493 (0.20)	7225 (0.15)
Total	47320 (1.00)	2480 (1.00)	49800 (1.00)

Table 2: Cluster assignments for test #2 (5000 contiguous records removed); column labels as in Table 1

Cluster	C-17 training	C-17 test	C-8, all data
1	12530 (0.28)	1409 (0.28)	14764 (0.30)
2	986 (0.02)	21 (0.00)	1484 (0.03)
3	5222 (0.12)	50 (0.01)	1975 (0.04)
4	4197 (0.09)	31 (0.01)	7828 (0.16)
5	5486 (0.12)	1122 (0.22)	7248 (0.15)
6	4642 (0.10)	831 (0.17)	7247 (0.15)
7	2991 (0.07)	406 (0.08)	1994 (0.04)
8	8745 (0.20)	1131 (0.23)	7260 (0.15)
Total	44799 (1.00)	5001 (1.00)	49800 (1.00)

Additional information about the effects of various data decimation schemes can be seen by normalizing each column by the total number of vectors used. When this is done for test case 1, classifications of the both the reference data and the test data are seen to faithfully replicate the class distributions derived from the training set. However, as might be expected, similar results are not achieved in cases 2 and 3, for which extended, contiguous data segments are removed. Despite the relatively large size of the reference data set, cases 2 and 3 removed sizeable fractions of the training sets (~10% and ~20%, respectively), and it is quite likely that these deleted sections contain unique scenes not found in the remaining data.

Table 3: Cluster assignments for test #3 (10000 contiguous records removed); column labels as in Table 1.

Cluster	C-17 training	C-17 test	C-8, all data
1	11676 (0.29)	1214 (0.12)	13199 (0.27)
2	909 (0.02)	34 (0.00)	1846 (0.04)
3	4011 (0.10)	501 (0.05)	7193 (0.14)
4	2848 (0.07)	757 (0.08)	3438 (0.07)
5	5211 (0.13)	1696 (0.17)	4616 (0.09)
6	6463 (0.16)	16 (0.00)	7093 (0.14)
7	3083 (0.08)	471 (0.05)	3976 (0.08)
8	5598 (0.14)	5312 (0.53)	8439 (0.17)
Total	44799 (1.00)	10001 (1.00)	49800 (1.00)

4. EXTENTION TO MAS SWATH

The analyses presented in the previous section verify the effectiveness and accuracy of classification schemes based on SOM-derived cluster centers. However, in those tests, a full-rank (i.e., 17-element) feature vector was always available in the classification phase. For the classifications attempted in this section, a full-rank feature vector will not be available; instead, only a subset of the original clustering space will be available for the classification task. In making classifications of this type, we are implicitly assuming that including the lidar data in the SOM analysis will significantly influence the characteristics of the MAS parameters that define the resulting cluster centers. However, the way in which these influences will manifest themselves in the MAS parameters is poorly understood, as is the magnitude of the changes they entail. We have therefore designed a simple empirical test with which to make an initial assessment of the validity of our classification scheme. We first classified each feature vector from the test sets using all of the 17 parameters that describe the cluster centers. After recording the cluster number assigned by this full-rank classification, we then reclassified the feature vector using only the 8 parameters derived from MAS. Once again, the cluster number assigned by this "rank-deficient" classification scheme was recorded.

Table 4: Percent of feature vectors for which the full rank and rank deficient classifications were identical

Test	Training Set	Reference Set	Test Data
1	68.5%	68.4%	68.1%
2	70.1%	70.3%	72.6%
3	74.1%	73.2%	69.7%

In Table 4 we list the percentage of classification matches for all feature vectors from each of the three

test cases. In our judgment, the success rate is quite good: the rank-deficient classification was identical to the full-rank classification in ~70% of all cases.

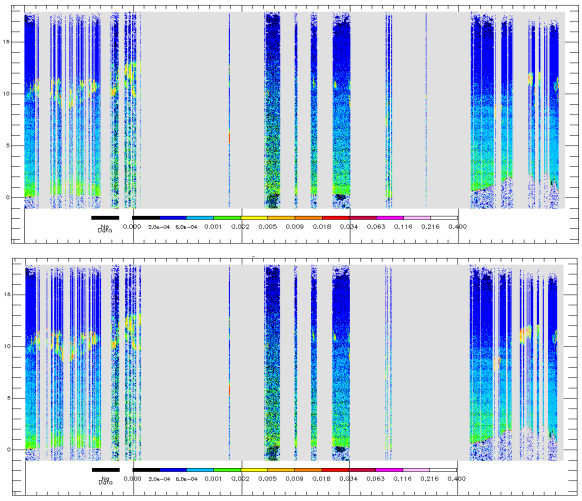

Fig. 2: CPL profiles from cluster 1. Upper panel: cluster membership when classified using the full MAS+CPL feature vector. Lower panel: class membership when analyzed using only the MAS elements of the feature vector.

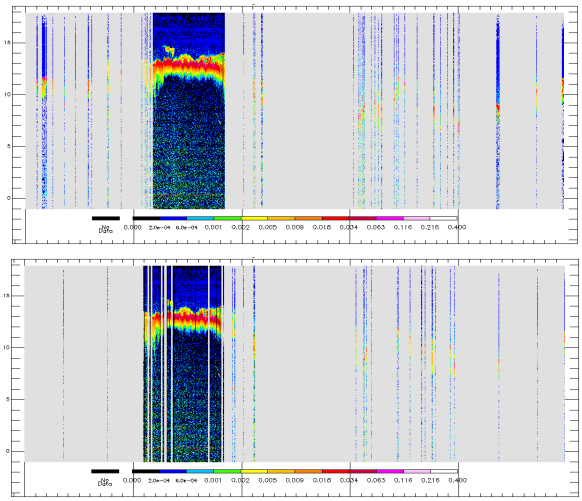

Fig. 3: CPL profiles from cluster 8. Upper panel: cluster membership when classified using the full MAS+CPL feature vector. Lower panel: class membership when analyzed using only the MAS elements of the feature vector.

The images shown in Fig. 2 and Fig. 3 provide a qualitative assessment of the difference between the two classification techniques for clusters 1 and 8, respectively. In both cases, the full-rank classification is shown in the upper panel, while the rank-deficient classification is shown in the lower panel. The visible differences between both image pairs are slight, though somewhat more noticeable for the dense cirrus cases characteristic of cluster 8.

The application of the rank-deficient nadir track classification scheme to the entire MAS swath is illustrated in Fig. 4. The data shown here is from the second leg of the November 24, 2002 flight.

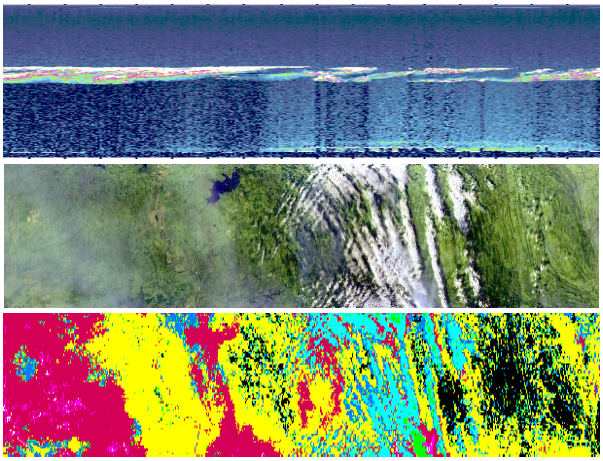

Fig. 4: TX-2002 measurements from 24 November 2002, flight leg 2. Upper panel: CPL attenuated backscatter coefficients at 532 nm showing an extended cirrus layer centered at ~11-km. Middle panel: MAS RGB image (channels 2, 10, and 20); the bright regions in the center right of the image are indicative of dense, low level clouds. Lower panel: full swath scene classification derived from MAS+CPL nadir-track SOM clusters

REFERENCES

1. T. Kohonen, *Self-Organization and Associative Memory, 3rd Edition*, Springer-Verlag, Berlin, 2000.
2. M. J. McGill, et al., The cloud physics lidar: Instrument description and initial measurement results, *Applied Optics*, Vol. 41, 3725– 3734, 2002.
3. King, M. D., et al., "Remote sensing of cloud, aerosol, and water vapor properties from the Moderate Resolution Imaging Spectrometer (MODIS)", *IEEE Trans. Geosci. Remote Sens.*, Vol. 30, 2–27, 1992.
4. Vaughan, M. A., et al., Multi-Sensor Data Fusion Technique: Part I, *Current Proceedings*
5. Vaughan, M. A., et al., Cloud and aerosol studies using combined CPL and MAS data, *Proc. SPIE*, Vol. 5571, 30–39, 2004.
6. D. M. Winker, et al., The CALIPSO mission: spaceborne lidar for observation of aerosols and clouds, *Proc. SPIE*, Vol. 4893, pp.1–11, 2003.
7. Rossow, W. B. and R. A. Schiffer, ISCPP Cloud Data Products, *Bull. Amer. Meteor. Soc.*, **72**, pp 2–20, 1991.