Advances in Remote Sensing, 2015, 4, 1-14
Published Online March 2015 in SciRes. http://www.scirp.org/journal/ars
http://dx.doi.org/10.4236/ars.2015.41001

Feature Analysis and Classification of Particle Data from Two-Dimensional Video Disdrometer

Sergey Gavrilov[1], Mamoru Kubo[2], Vu Anh Tran[1], Duc Luu Ngo[1], Ngoc Giang Nguyen[1], Lan Anh T. Nguyen[1,3], Favorisen Rosyking Lumbanraja[1], Dau Phan[1], Kenji Satou[2]

[1]Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
[2]Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
[3]Department of Computer Science, Hue University of Education, Hue, Vietnam
Email: gavriloff.sv@gmail.com, kubom@se.kanazawa-u.ac.jp, tvatva2002@gmail.com, ndluu@blu.edu.vn, giangnn.bkace@gmail.com, lananh257@gmail.com, favorisen@gmail.com, pdaukg@gmai.com, ken@t.kanazawa-u.ac.jp

Received 9 January 2015; accepted 26 January 2015; published 30 January 2015

Abstract

We developed a ground observation system for solid precipitation using two-dimensional video disdrometer (2DVD). Among 16,010 particles observed by the system, around 10% of them were randomly sampled and manually classified into five classes which are snowflake, snowflake-like, intermediate, graupel-like, and graupel. At first, each particle was represented as a vector of 72 features containing fractal dimension and box-count to represent the complexity of particle shape. Feature analysis on the dataset clarified the importance of fractal dimension and box-count features for characterizing particles varying from snowflakes to graupels. On the other hand, performance evaluation of two-class classification by Support Vector Machine (SVM) was conducted. The experimental results revealed that, by selecting only 10 features out of 72, the average accuracy of classifying particles into snowflakes and graupels could reach around 95.4%, which had not been achieved by previous studies.

Keywords

Solid Precipitation, Particle Classification, 2DVD, Fractal Dimension, PCA

1. Introduction

Due to the diversity of terrain, rainfall and snowfall phenomena take on different forms depending on location.

How to cite this paper: Gavrilov, S., Kubo, M., Tran, V.A., Ngo, D.L., Nguyen, N.G., Nguyen, L.A.T., Lumbanraja, F.R., Phan, D. and Satou, K. (2015) Feature Analysis and Classification of Particle Data from Two-Dimensional Video Disdrometer. *Advances in Remote Sensing*, **4**, 1-14. http://dx.doi.org/10.4236/ars.2015.41001

For instance, in the coastal area facing the Sea of Japan, snow clouds develop over the sea by the winter monsoon wind blowing from the north-west. Kanazawa is a coastal city that experiences rather heavy snowfalls despite being located at a low latitude (approx. 36˚N) [1] [2]. Also in this place, amount and type of precipitation change quite rapidly. Therefore, it is important to monitor them continuously for decreasing the damage of heavy snowfall as well as meteorological understanding of orographic snowfall. Especially, it is important to understand the snowfall formation mechanism with different types of solid precipitation such as snowflake and graupel.

A polarimetric radar [3]-[6] is a popular facility for measuring precipitation intensity in wide area. Additionally, a disdrometer is used for the ground-based observation of precipitation at a spot. It is a relatively-small instrument which can measure the size and falling velocity of a particle. Based on the fact that rain and graupel have different distribution of size and falling velocity, it is possible to discriminate them using a disdrometer. However, if two particles have similar size and falling velocity, it is impossible to discriminate them by a disdrometer. In this sense, the observation of precipitation using a polarimetric radar and/or a disdrometer is not sufficient for accurately estimating the amount of precipitation consisting of various types.

In addition to size and falling velocity, the shape of each particle is significantly useful for the classification of particle types. In our previous study, we proposed a system for automatically taking particle images by a CCD video camera and classifying them into snowflakes and graupels [7]. Using rich information contained in a large number of grayscale particle images, the system achieved high accuracy of classification (94.14%). However, it is not an easy-to-purchase product and requires large space comparable with a room. One more disadvantage of this system is that it utilizes only the combination of basic classifiers with only one best pair of features from more than ten available features.

As a more popular instrument, a two-dimensional video disdrometer (2DVD) has been used for characterizing solid precipitation at the ground [8]. The instrument is manufactured by Joanneum Research of Austria. 2DVD measures volume, diameter, shape, and velocity of every individual particle. From these data, one can estimate particle size distribution, precipitation rate, and other related variables. As to the classification, recently a hydrometeor classification system with 2DVD has been proposed [9]. However, since it can only give a dominant type of precipitation observed in a time interval (60 sec.), it is not available for the purpose of particle-by-particle classification indispensable in accurately estimating the amount of mixed-type precipitation.

In this study, we developed a new system with 2DVD for observing and estimating various particles. Although the 2DVD takes binary image with lower resolution than CCD video camera, combination of up-to-date classifier and features including fractal-related ones enables the system to outperform the accuracy achieved in our previous study.

Rest of this paper is organized as follows. Section 2 gives descriptions about the 2DVD system, weather condition of observation, representation of observed particles as feature vectors, and various machine learning algorithms used in this study. In Section 3, the results of feature analysis are shown and followed by the results of performance evaluation on the proposed system for classifying snowflakes and graupels. Finally, Section 4 concludes this paper.

2. Materials and Methods

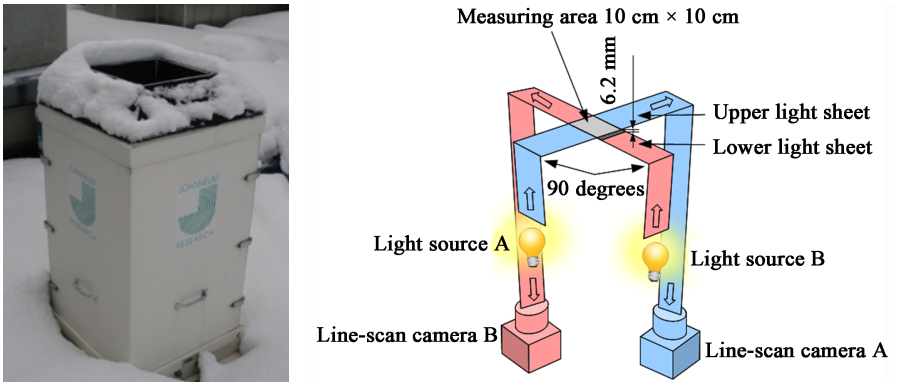
2.1. System and Condition of Observation

2DVD is an optical device developed for measuring precipitation drop size, shape, and velocity field. **Figure 1** shows the 2DVD sensor unit. The sensor unit consists of two orthogonal and synchronized line-scan cameras and a bright light source in front of each of them. While precipitation particles fall between the cameras and light sources (an area of 10 cm × 10 cm) their shapes are recorded as shadows are being projected. We have observed snowfall event from 1250 JST to 1300 JST in January 26, 2011 at Kanazawa University. The data of 16,010 snow particles were recorded by the 2DVD. **Figure 2** shows MTSAT-2 satellite image at 1200 JST 26 January 2011 and the location of observation point. The air temperature was about 0˚C through the event duration.

2.2. Preparation of Data for Analysis and Classification

2.2.1. Particle Images and Basic Features

Figure 3 illustrates examples of particle image data recognized and generated by 2DVD. Since 2DVD scans two



(a) (b)

Figure 1. 2DVD sensor unit. (a) Photograph of the sensor unit covered with snow; (b) Illustration of sensor unit construction.

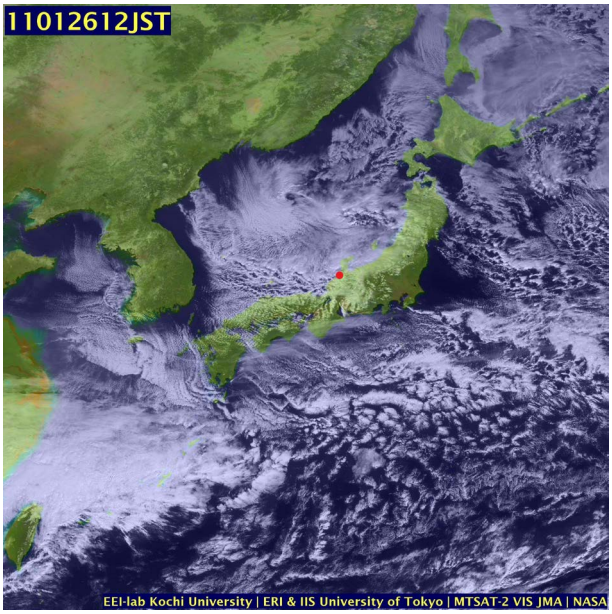


Figure 2. MTSAT-2 satellite image at 1200 JST 26 January 2011 (from http://weather.is.kochi-u.ac.jp/). The 2DVD is installed at Kanazawa University and the location of observation point is indicated by a red circle. (36.544˚N, 136.705˚E).

line images at once from two orthogonally oriented cameras (A and B), two different images are obtained for each particle.

In **Figure 3**, it can be seen that a graupel is round-shaped as an approximate ellipse, and in contrast, a snowflake has a complex shape. As to the size of a particle, graupels are relatively smaller than snowflakes. These features meet intuitive criteria in human's discrimination of snowflake and graupel. The latter feature was frequently used in previous studies since it is easier to observe.

In addition to shape and size, it is possible to obtain various features of a particle by using 2DVD. The list of features used in this study is shown in **Table 1**.

The 2DVD software computes the volume and equivolumetric diameter based on three-dimensional shape reconstructed from two orthogonal projections. The particle shadows in the upper light sheet are matched with particle shadows in the lower sheet, and the software obtains the vertical fall velocity and height quantization (height_of_one_line) from the falling time through the planes separated 6.2 mm vertically at the line-scan rate of 34.1 kHz. The number of lines scanned by each camera is the height of the particle. The light sheet of 10 cm is

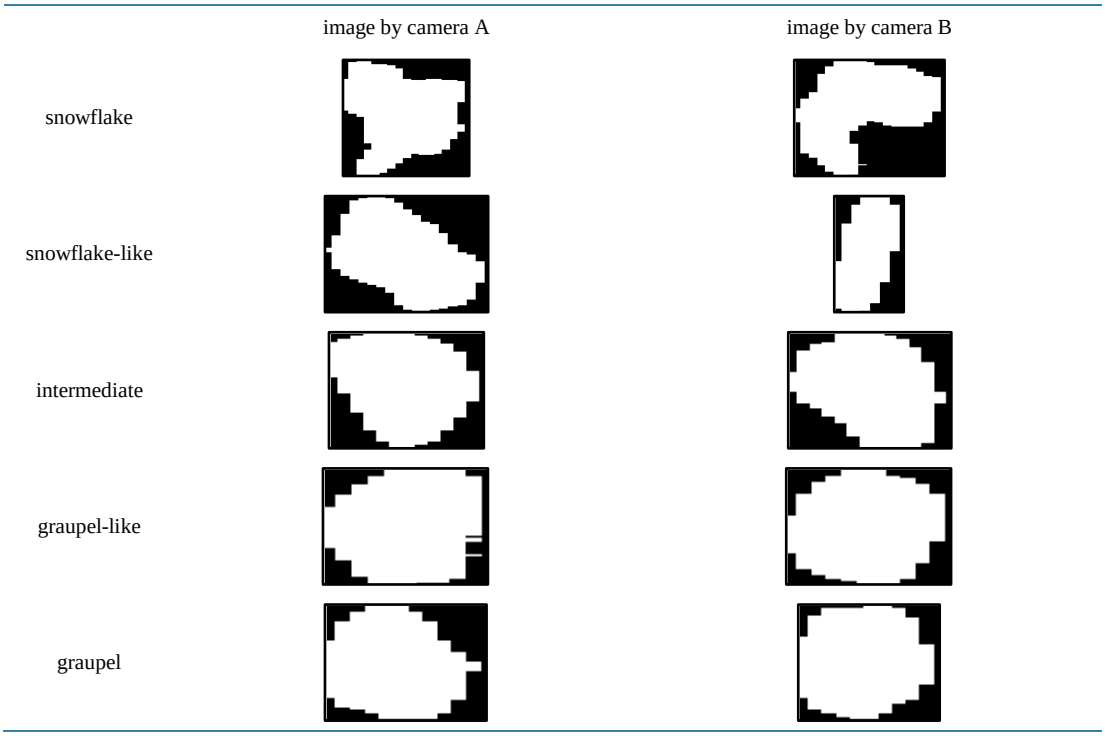


Figure 3. Particle images taken by 2DVD.

Table 1. Features for analysis and classification.

Feature type	Feature name
Camera-independent features	Equivolumetric_diameter[mm], volume[mm^3], vertical_fall_velocity[m/s], height_of_one_line[mm]
Camera-specific features	Height [mm]_A, height [mm]_B, number_of_lines_A, number_of_lines_B, pixelwidth[mm]_A, pixelwidth [mm]_B, width [pixel]_A, width [pixel]_B, height [pixel]_A, height[pixel]_B, total_pixels_A, total_pixels_B, area [mm^2]_A, area [mm^2]_B, perimeter [mm]_A, perimeter[mm]_B, box_count_1_A, box_count_1_B, box_count_2_A, box_count_2_B, box_count_4_A, box_count_4_B, box_count_8_A, box_count_8_B, fractal_1_2_A, fractal_1_2_B, fractal_2_4_A, fractal_2_4_B, fractal_1_4_A, fractal_1_4_B, fractal_4_8_A, fractal_4_8_B, fractal_2_8_A, fractal_2_8_B
Camera-independent features (max and min) converted from camera-specific features (A and B)	Height [mm]_max, height [mm]_min, number_of_lines_max, number_of_lines_min, pixelwidth [mm]_max, pixelwidth [mm]_min, width [pixel]_max, width [pixel]_min, height [pixel]_max, height [pixel]_min, total_pixels_max, total_pixels_min, area [mm^2]_max, area [mm^2]_min, perimeter [mm]_max, perimeter [mm]_min, box_count_1_max, box_count_1_min, box_count_2_max, box_count_2_min, box_count_4_max, box_count_4_min, box_count_8_max, box_count_8_min, fractal_1_2_max, fractal_1_2_min, fractal_2_4_max, fractal_2_4_min, fractal_1_4_max, fractal_1_4_min, fractal_4_8_max, fractal_4_8_min, fractal_2_8_max, fractal_2_8_min
Other features (not used in analysis and classification)	Time

mapped onto 512 pixels in the line-scan camera, and the horizontal resolution of pixel (pixel width) is about 0.2 mm. The longest scan line is the particle width. The area of each particle was computed by multiplying total number of pixels (total_pixels), height_of_one_line and pixel width. We got the boundary of particle shape and computed the particle perimeter.

Camera-specific features are important since they contain various information obtained by 2DVD. However, it is not sufficient to use them directly in the analysis and classification. When we use machine learning algorithms listed in subsection 2.3, the same type of features obtained by cameras A and B (e.g. perimeter [mm]_A and perimeter[mm]_B) are also treated as simply different and independent ones. To overcome this problem, we

added extra features that are the result of integrating camera-specific features by calculating maximum and minimum values (**Figure 4**). For example, if perimeter [mm]_A > perimeter [mm]_B, then perimeter[mm]_max = perimeter [mm]_A and perimeter[mm]_min = perimeter[mm]_B. In a sense, it is a sorting operation of values from two cameras and if a feature is mainly characterized by large (small) values of it, the integrated feature of its maximum (minimum) will have strong power in the analysis and classification of particles.

2.2.2. Fractal-related Features

Perimeter is a feature that reflects two different characteristics of particle, that is, size and complexity of shape. In this study, we introduced fractal-related features also related to complexity of shape.

Fractal geometry provides a mathematical model for many complex objects with property of self-similarity found in nature. Fractal dimension is a useful feature for shape classification. The snowflake formation modeled by fractal dimension, was proposed for improvement estimates of snowfall retrieval by radar remote sensing [10] [11]. This study uses the box-counting method, which is one of the frequently used techniques to estimate the fractal dimension also known as Minkowski dimension [12] [13]. First, the smallest number of box shaped elements covering the particle boundary is counted (**Figure 5**). Next, the obtained amount of covering elements is log-log plotted versus the reciprocal of the element size (**Figure 6**). Finally, the box dimension estimate is taken from the monotonically rising linear slope.

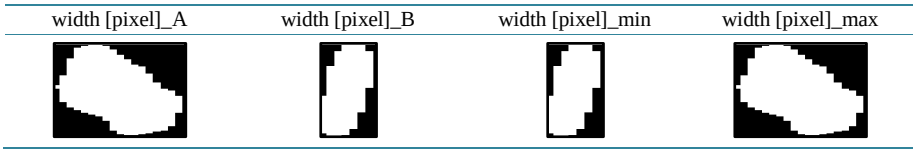


Figure 4. Integration of camera-specific features into max and min values.

Figure 5. Example of covering results from the box-counting method. (a) Snowflake by camera A; raw image by 2DVD (leftmost), boundary covered by boxes of size 1, 2, 4, and 8; (b) Snowflake by camera B.

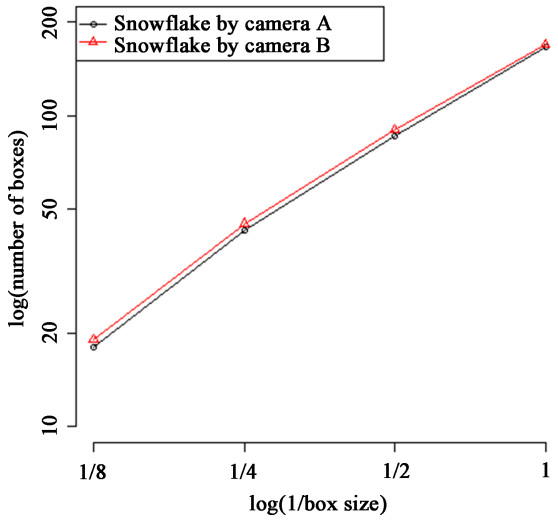


Figure 6. The log-log plot of the box-counting method.

2.2.3. Human Annotation

Total number of particles in our dataset is 16,010, that is, it consists of 16,010 feature vectors with the features listed in **Table 1**. To conduct meaningful analysis and evaluation of classification performance, we randomly sampled 1600 feature vectors and annotated them manually. Before annotation, five categories were prepared: *snowflake*, *snowflake-like*, *intermediate*, *graupel-like*, and *graupel*. Additionally, if one of two images for a particle matched one of the following rules, it was automatically annotated as *warning* and filtered out before random sampling since it can be regarded as outlier or erroneous data.

- equivolumetric_diameter [mm] is less than 0.2.
- vertical_fall_velocity [m/s] is greater than 4.
- Width [pixel]/height [pixel] is less than 1/3 or greater than 3.
- The horizontal position of the particle in the raw image is left-end and over 50% of left edge of the particle image is occupied by black pixel (*i.e.* it is strongly suspected that the particle passed by the left end of a camera and whole image of it was not taken by 2DVD).

The numbers of annotated samples are shown in **Table 2**. According to these annotations, the datasets shown in **Table 3** are used for analysis and classification in Section 3.

2.3. Algorithms

In this subsection, the algorithms we used for analysis and classification are being described.

2.3.1. Normalization

A feature vector consists of two or more feature values for features. However, it is problematic to use the original values for machine learning because in general, value distribution can differ from feature to feature. Therefore, it is popular to normalize the original values of feature vectors so that all the features have the same average and variance. In this study, we normalized our dataset with average = 0 and variance = 1 for each feature before the analysis and classification.

2.3.2. Pearson's Correlation Coefficient

To see the direct and pairwise relationship between every pair of features, we calculated Pearson's correlation

Table 2. The number of samples after annotation.

Annotation	The number of particles
Snowflake	559
Snowflake-like	111
Intermediate	39
Graupel-like	144
Graupel	747
Warning	2,118
Not annotated	12,292

Table 3. Datasets according to annotation.

Dataset	Annotation	The number of particles
Whole	Snowflake, snowflake-like, intermediate, graupel-like, graupel, warning, not annotated	16,010
No-warning	Snowflake, snowflake-like, intermediate, graupel-like, graupel, not annotated	13,892
Warning-only	Warning	2118
5-Classes	Snowflake, snowflake-like, intermediate, graupel-like, graupel	1600
2-Classes	Snowflake, graupel	1306

coefficient. If its value is near to 1, two features are quite similar. It is one of the most basic feature analysis methods. In addition, it is known that, removing one of two similar and redundant features may lead to better performance of classification, regression, clustering, and other machine learning tasks.

2.3.3. Principal Component Analysis (PCA)

Among various unsupervised learning algorithms, PCA might be the most popular one. Based on the calculation of features' linear combination that maximizes the variance, PCA converts the original feature space into the space of principal components (PCs). After PCA, all the PCs are ordered as PC1, PC2, … and it is believed that PC1 is the strongest feature for characterizing the feature vectors, PC2 is secondly strong, and so on. Due to this effect of PCA, it is broadly used for different purposes. As the basic analysis of original features, coefficient of each feature in the linear combination formula for some important PCs like PC1 is evaluated. In this study, it may reflect the importance of the feature to characterize and classify snowflakes and graupels.

2.3.4. Support Vector Machine (SVM)

SVM was first developed by Vladimir Vapnik [14]. Due to its applicability and high-performance, it is one of the most popular machine learning algorithms today. Among various variants and implementations of SVM, we used ksvm function implemented in kernlab package for R. Regarding the choice of kernel, the default one (Radial Basis Function kernel, also known as Gaussian kernel) was adopted. A hyper-parameter "sigma" for this kernel is being automatically optimized by ksvm.

2.3.5. Cross-Validation

To evaluate the performance of predicting the class label (*i.e.* snowflake or graupel) of unseen samples (*i.e.* unseen particles), it is popular to conduct cross-validation. In this study, we adopted 10-fold cross-validation that randomly divides given dataset into 10 and perform learning and prediction 10 times by changing 10% of dataset for test (rest of 90% is used for training). One problem about this kind of cross-validation is that the evaluated performance is affected by the result of random division and different performances are achieved in every evaluation. To solve this problem, we repeated 10-fold cross-validation 100 times and averaged the accuracy.

3. Experimental Results and Discussion

3.1. Feature Analysis by Pearson's Correlation Coefficient

Figure 7 illustrates the result of correlation analysis on all feature pairs. It can be summarized as follows:

- Box-count features (*i.e.* features about the number of boxes) are highly similar to each other. In contrast, fractal features are dissimilar to each other.
- Some of other features are similar to each other (*i.e.* height and perimeter features). It indicates that redundant features like box-count may exist also in these other features.

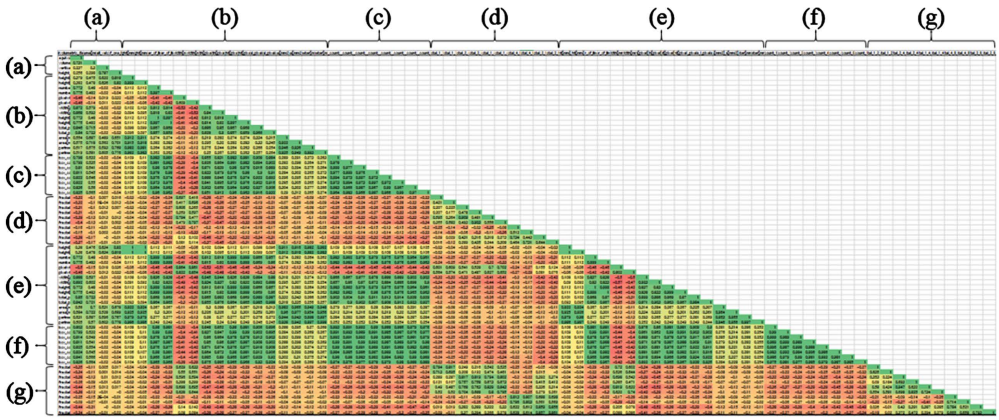


Figure 7. Correlation analysis of features. Green (red) color corresponds to high (low) value. (a) Camera-independent features; (b) Camera-specific features; (c) box-count features; (d) fractal features; (e) (f) (g) Max and min of (b) (c) (d).

- About the difference between camera-specific features ((b), (c), and (d)) and camera-independent features ((e), (f), and (g)) calculated from them, fractal features (d) and (g) showed clear difference. In other words, calculation of max and min was meaningful at least for fractals.

3.2. Feature Analysis by PCA

Figures 8-10 illustrate the similarity among the principal components 1-3 in four datasets (except "warning-only"). In each figure, features are sorted in descending order of principal component of whole dataset. Top 10 important features in each dataset and PC are shown in **Tables 4-6**. From these figures and tables, it can be clearly seen that:

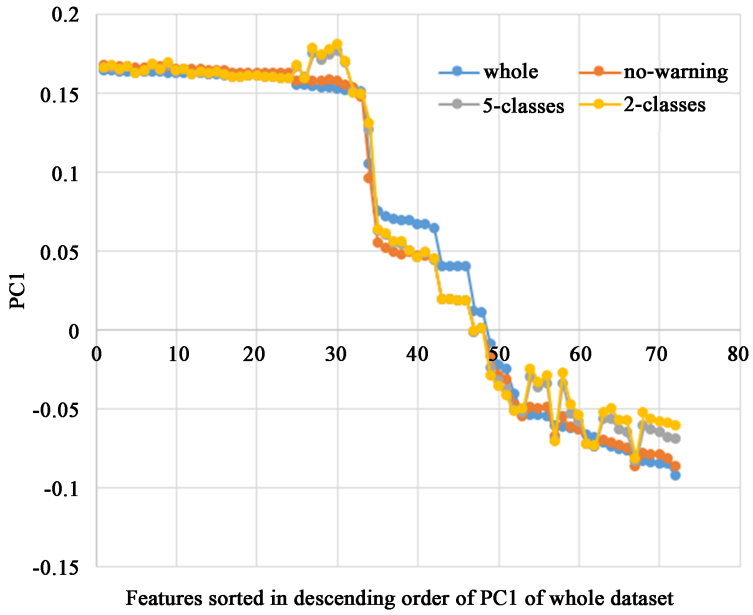


Figure 8. PC1 of the datasets except "warning-only".

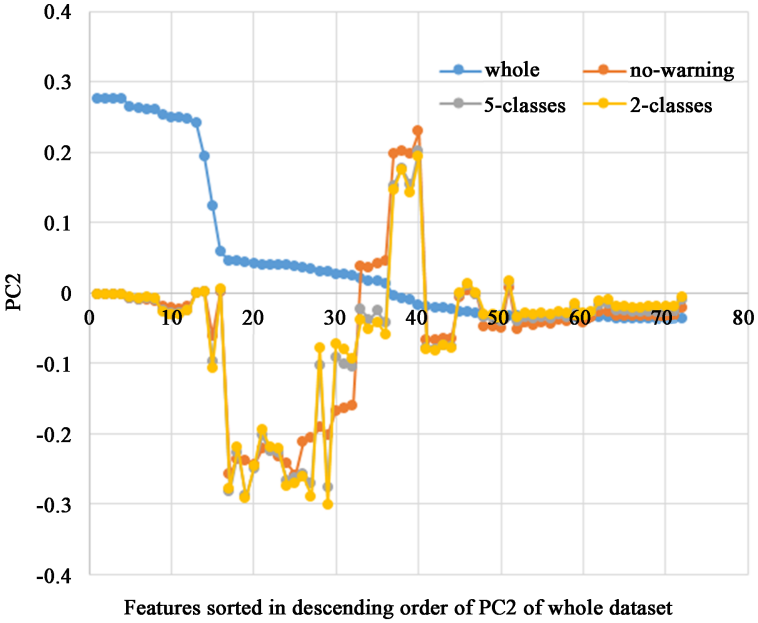


Figure 9. PC2 of the datasets except "warning-only".

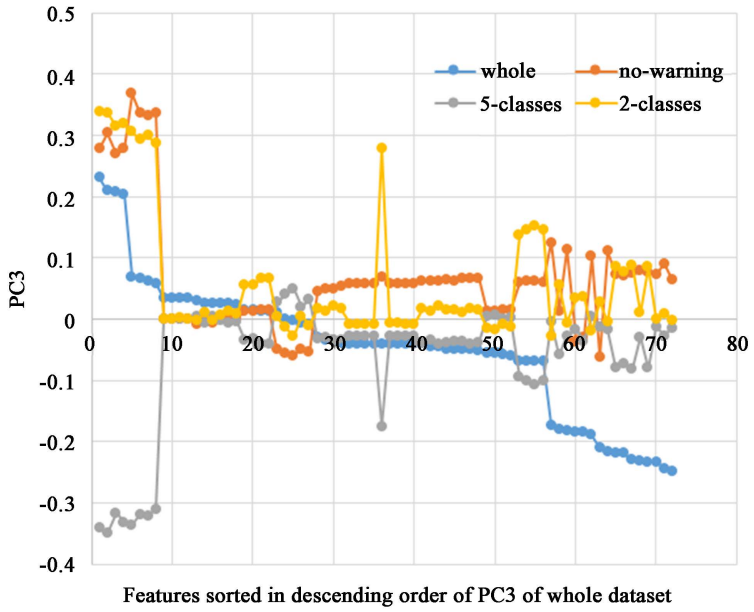


Figure 10. PC3 of the datasets except "warning-only".

Table 4. Top 10 features in descending order of PC1 values.

rank	whole	no-warning	5-classes	2-classes	warning-only
1	box_count_4_min	box_count_4_min	total_pixels_B	total_pixels_B	height[mm]_min
2	box_count_8_max	box_count_8_min	total_pixels_max	total_pixels_max	height[mm]_B
3	box_count_4_max	box_count_8_max	total_pixels_min	total_pixels_min	height[mm]_max
4	box_count_4_B	box_count_8_B	total_pixels_A	total_pixels_A	height[mm]_A
5	box_count_4_A	box_count_4_B	width[pixel]_B	width[pixel]_B	perimeter[mm]_min
6	box_count_2_min	box_count_4_max	box_count_8_B	box_count_8_B	perimeter[mm]_B
7	box_count_8_min	box_count_8_A	box_count_8_min	box_count_8_min	perimeter[mm]_A
8	box_count_8_A	box_count_2_min	box_count_4_B	width[pixel]_max	perimeter[mm]_max
9	box_count_8_B	box_count_4_A	width[pixel]_max	box_count_8_max	area[mm2]_max
10	box_count_2_max	box_count_2_B	box_count_8_max	box_count_4_B	area[mm2]_min

Table 5. Top 10 features in descending order of PC2 values.

rank	whole	no-warning	5-classes	2-classes	warning-only
1	height[mm]_B	fractal_4_8_min	fractal_4_8_min	fractal_4_8_min	pixelwidth[mm]_min
2	height[mm]_min	fractal_4_8_B	fractal_4_8_B	fractal_4_8_B	pixelwidth[mm]_B
3	height[mm]_max	fractal_4_8_A	fractal_4_8_A	fractal_4_8_max	pixelwidth[mm]_max
4	height[mm]_A	fractal_4_8_max	fractal_4_8_max	fractal_4_8_A	pixelwidth[mm]_A
5	perimeter[mm]_min	fractal_2_8_min	width[pixel]_min	width[pixel]_min	fractal_1_2_max
6	perimeter[mm]_B	fractal_2_8_B	width[pixel]_A	width[pixel]_A	fractal_1_2_min
7	perimeter[mm]_A	fractal_2_8_max	equivolumetric_diameter[mm]	equivolumetric_diameter[mm]	fractal_1_2_B
8	perimeter[mm]_max	fractal_2_8_A	vertical_fall_velocity[m/s]	vertical_fall_velocity[m/s]	fractal_1_2_A
9	area[mm2]_max	width[pixel]_min	height_of_one_line[mm]	width[pixel]_B	height_of_one_line[mm]
10	area[mm2]_B	width[pixel]_A	width[pixel]_B	height_of_one_line[mm]	fractal_1_4_max

Table 6. Top 10 features in descending order of PC3 values.

rank	whole	no-warning	5-classes	2-classes	warning-only
1	fractal_4_8_min	fractal_2_8_min	width[pixel]_A	fractal_4_8_min	pixelwidth[mm]_min
2	fractal_4_8_max	fractal_2_8_A	width[pixel]_min	fractal_4_8_max	pixelwidth[mm]_B
3	fractal_4_8_B	fractal_2_8_B	width[pixel]_max	fractal_4_8_A	pixelwidth[mm]_A
4	fractal_4_8_A	fractal_2_8_max	equivolumetric_diameter[mm]	fractal_4_8_B	volume[mm3]
5	fractal_2_8_min	fractal_4_8_max	width[pixel]_B	fractal_2_8_min	width[pixel]_max
6	fractal_2_8_B	fractal_4_8_min	box_count_8_A	fractal_2_8_max	width[pixel]_A
7	fractal_2_8_max	fractal_4_8_A	height_of_one_line[mm]	fractal_2_8_B	box_count_8_max
8	fractal_2_8_A	fractal_4_8_B	vertical_fall_velocity[m/s]	fractal_2_8_A	box_count_8_A
9	height[mm]_max	fractal_2_4_min	box_count_8_max	volume[mm3]	total_pixels_max
10	height[mm]_A	fractal_2_4_B	fractal_2_4_A	total_pixels_B	box_count_8_B

- PC1s of these datasets are similar to each other (**Figure 8**). Most of the important features in PC1 are occupied by box-count features (**Table 4**).
- PC2 of the dataset "whole" is quite dissimilar to others (**Figure 9**) and the difference is caused by the inclusion of "warning-only". In other words, after filtering errors, PC2 is more or less the same in each dataset. About top 10 features of PC1 of "warning-only" (**Table 4**), it is convincing that most of them are occupied by size-related features (height, perimeter, area, etc.) because many of the particles in this dataset were removed from "whole" dataset due to their strange size. About PC2s of the datasets "no-warning", "5-classes", and "2-classes", some of the fractal features occupy top 4 important features (**Table 5**).
- In **Figure 10**, PC3s of the datasets "5-classes" and "2-classes" are quite dissimilar (correlation between them is −0.97). Since in "2-classes", ambiguous particles annotated as "snowflake-like", "intermediate", or "graupel-like" are removed from "5-classes", it can be interpreted that PC3 of "5-classes" is highly affected by the characteristics of such ambiguous particles.

For visually understanding the sample distribution, we show 3D plots of the datasets. In **Figure 11**, it can be seen that the distributions of samples in three datasets "no-warning", "5-classes", and "2-classes" are almost the same. The 3D plots from three angles for "5-classes" show that, snowflake samples have their own distribution distinguishable from others. In contrast, samples of other annotations (snowflake-like, intermediate, graupel-like, and graupel) are distributed in the plane near to the PC2-PC3. About the L-like distribution of these samples, it is caused by the combined use of camera-specific fractal features (fractal_1_2_A, …, fractal_2_8_B) and camera-independent fractal features (fractal_1_2_max, …, fractal_2_8_min). For example, removal of box-count features does not affect to the L-like shape of the distribution, however, removal of camera-specific or camera-independent fractal features makes it ambiguous (**Figure 12**). Although the meaning of the distribution is still unclear, this result suggests that the fractal features could provide more detailed classification of non-snowflake particles.

3.3. Particle Classification by SVM

As shown in **Table 1**, 72 features are available for training a statistical model to classify given samples (particles) into snowflakes and graupels. Using the algorithms described in subsections 2.3.4 and 2.3.5, first we evaluated the accuracy of prediction with "2-classes" dataset and all 72 features. The average error of prediction (*i.e.* 1 - average accuracy) was 0.08263. After converting the 72 features into 72 PCs by PCA, the average error decreased to 0.07191.

Since so many redundant features exist in the 72 features, reduction of feature set by feature selection might decrease the average error of prediction. Although various algorithms have been proposed for fully-automatic feature selection, in this study we initially tried to select a representative feature in each feature group, assuming a feature group consisting of all features with common name prefix. For example, perimeter [mm]_A, perimeter[mm]_B, perimeter[mm]_max, and perimeter[mm]_min belong to the same group. In case of box-count and

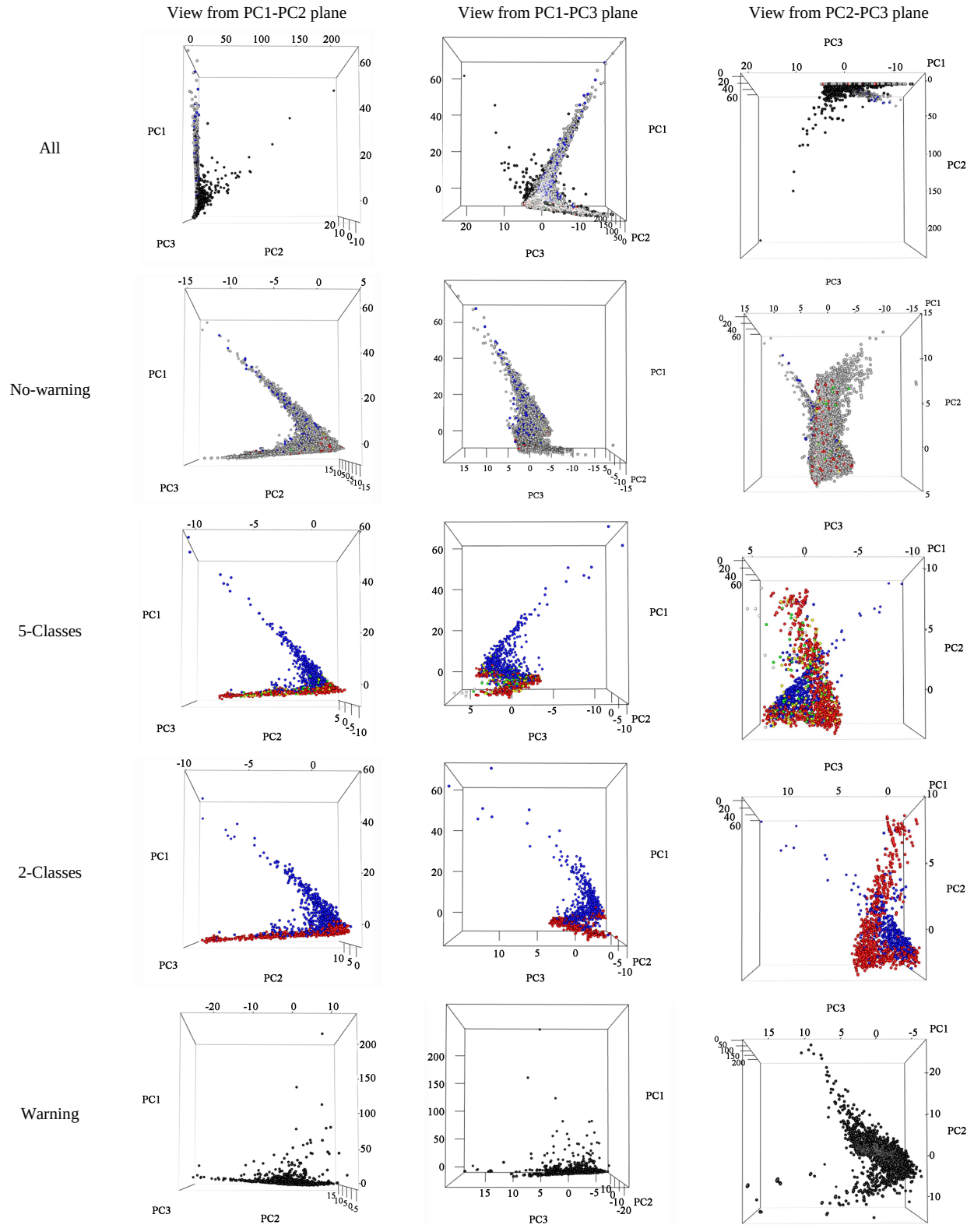


Figure 11. 3D plots of PC1, PC2, and PC3 in five datasets from three different angles of view. The colors of points (gray, black, blue, green, white, yellow, red) indicate the annotations (not annotated, warning, snowflake, snowflake-like, intermediate, graupel-like, graupel), respectively.

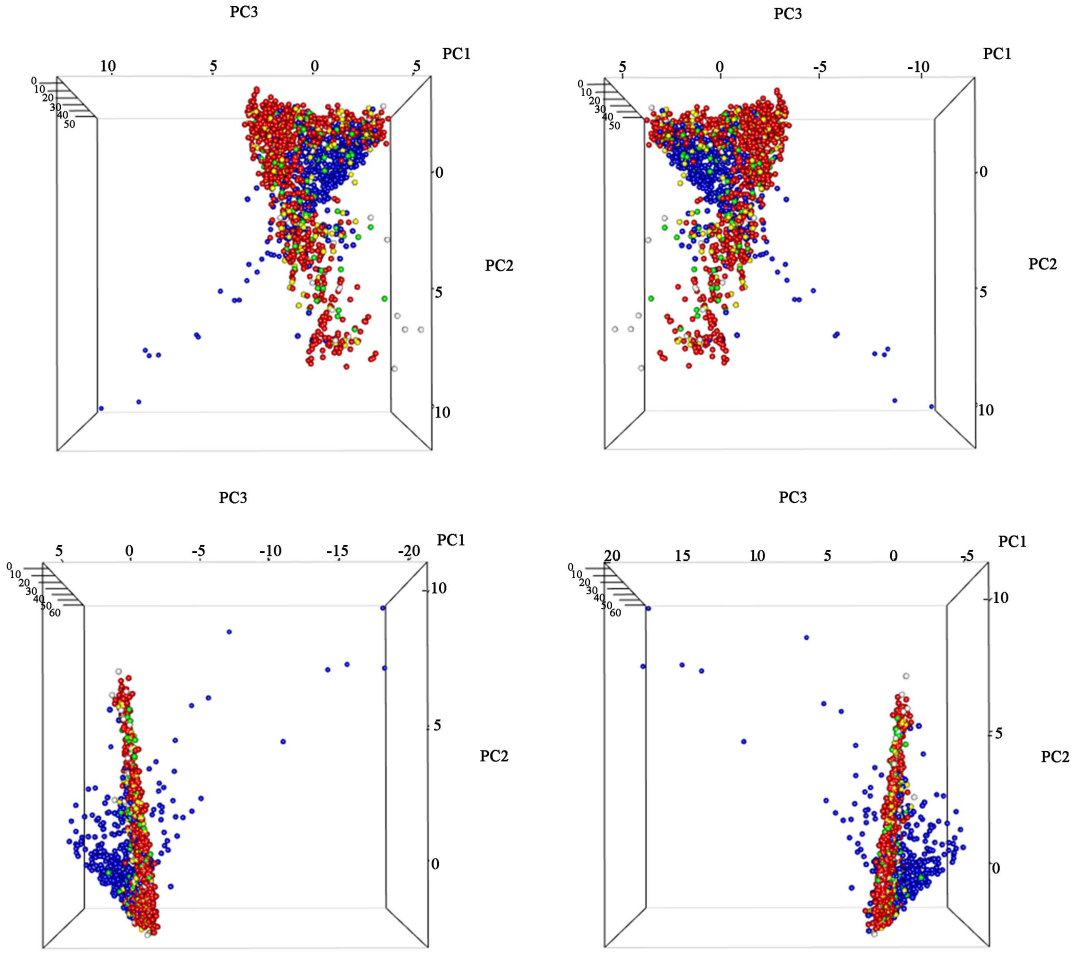


Figure 12. 3D plots of the dataset "5-classes" without some features. In top-left and top-right panels, camera-specific and camera-independent box-count features are removed, respectively. Also in bottom-left and bottom-right panels, camera-specific and camera-independent fractal features are removed, respectively.

fractal features, numbers in the names were ignored since they are homogenous except the parameters for calculating them. To choose the representative feature in each group, 72 evaluations were performed using only one specific feature in each evaluation. As a result, 14 representative features with the lowest average errors in their groups were selected (**Table 7**). Among them, box_count_2_max achieved the best performance (0.1055) as a single feature. It is also notable that the suffixes "_max" and "_min" frequently appear instead of "_A" and "_B". It indicates that the conversion of camera specific features to camera-independent ones contributed to achieve better classification performance.

Starting from the feature set with all of these 14 features, feature selection by backward elimination was performed. It is an iterative feature selection method which removes a feature in an iteration. If the size of feature set in the iteration i is n_i, all subsets with size $n_i - 1$ are evaluated, and if the elimination of a feature achieved the best improvement of average error, it is removed in the next iteration. As a baseline performance before the 1st iteration, the average error 0.0543 achieved by the feature set with all of these 14 features was used.

In this study, four features were removed through 1st to 4th iterations, and the process of backward elimination stopped since 5th iteration could not achieve any improvement. Using the remaining 10 features, the average error 0.0461 was achieved and it was the best performance of classification in this study[1]. Unlike the analysis in subsection 3.2, this result revealed that fractal features could not contribute to the best performance. In other

[1]We conducted t-test on two groups of errors before calculating 0.0465 and 0.0461 in **Table 7**, but it did not show statistically significant difference (p-value = 0.05153). However, at least it was confirmed that 0.0484 and 0.0465 were significantly different (p-value = 3.815e−14).

Table 7. Average errors (*i.e.* 1-average accuracy) in the predictions by single feature and multiple features with backward elimination. Before backward elimination, average error of prediction by using all 14 features listed in the first column was 0.0543. In each iteration of backward elimination, if the elimination of a feature decreased (increased) the average error of prediction, it is shown in red (blue) color. The least average error in each column is shown in bold face and the corresponding feature is being not used in the succeeding iterations of backward elimination.

feature	prediction by single feature	1st iteration	2nd iteration	3rd iteration	4th iteration	5th iteration
box_count_2_max	0.1055	0.0599	0.0543	0.0481	0.0493	0.0463
total_pixels_max	0.1198	0.0577	0.0538	0.0485	**0.0461**	removed
number_of_lines_min	0.1222	0.0549	0.0511	0.0485	0.0480	0.0466
height[pixel]_min	0.1224	0.0548	0.0513	0.0481	0.0480	0.0467
perimeter[mm]_max	0.1274	0.0683	0.0665	0.0626	0.0654	0.0653
width[pixel]_max	0.1405	0.0564	0.0509	0.0471	0.0479	0.0476
area[mm2]_max	0.1886	0.0602	0.0574	0.0495	0.0526	0.0522
height[mm]_min	0.1913	0.0546	0.0531	**0.0465**	removed	removed
equivolumetric_diameter [mm]	0.2026	0.0652	0.0622	0.0556	0.0561	0.0573
volume[mm3]	0.2045	0.0567	0.0506	0.0481	0.0486	0.0469
fractal_2_8_min	0.2069	0.0520	**0.0484**	removed	removed	removed
pixelwidth[mm]_max	0.2434	**0.0517**	removed	removed	removed	removed
height_of_one_line [mm]	0.3449	0.0557	0.0529	0.0509	0.0504	0.0513
vertical_fall_velocity [m/s]	0.4261	0.0556	0.0522	0.0503	0.0499	0.0503

words, they might be useful for more detailed characterization of various particles, not for just classifying snowflakes and graupels. In contrast, a box-count feature (box_count_2_max) was so important as to the classification by only one feature achieved average error 0.1055 that is nearly 90% accuracy. It is an interesting finding that, although a box-count feature is a by-product of fractal calculation, it is significantly important in the classification of snowflakes and graupels.

4. Conclusions

In this study, we conducted feature analysis and classification of particle data from 2DVD through the combined use of various statistical methods including supervised and unsupervised machine learning. Experimental results revealed that fractal and box-count features were useful for the characterization and classification of snowflakes and graupels. The average accuracy of particle-by-particle classification was around 95.4%, which had not been achieved by previous studies. From this result, it could be said that we could develop a system for automatic solid precipitation monitoring with practically sufficient accuracy of discriminating snowflakes and graupels.

In **Table 1**, we mentioned that each particle datum was attached to its timestamp of observation. Combining time information with the results of classification on large amount of particles, it was possible to conduct time-series analysis of amount and type of particles, which contributed to elucidate the mechanism of orographic snowfall (phenomena). Furthermore, conducting human annotation with not only two types (*i.e.* snowflake and graupel) but also other detailed types of particles (e.g. dendrite-like, aggregate-like, melting-snow-like, etc.), it was becoming possible to quantitatively analyze wide-variety of snowfall in places with weather conditions similar to Kanazawa. The ground-based measurements of snow particles and identification of snow type would be useful for deriving radar reflectivity-snow rate relationships.

Acknowledgements

The authors would like to thank President Ken-Ichiro Muramoto of Ishikawa National College of Technology and Professor Yasushi Fujiyoshi of Hokkaido University for cooperation in the snowfall observation.

References

[1] Ohigashi, T. and Tsuboki, K. (2005) Structure and Maintenance Process of Stationary Double Snowbands along the Coastal Region. *Journal of the Meteorological Society of Japan*, **83**, 331-349. http://dx.doi.org/10.2151/jmsj.83.331

[2] Harimaya, T., Kodama, H. and Muramoto, K. (2004) Regional Differences in Snowflake Size Distributions. *Journal of the Meteorological Society of Japan*, **82**, 895-903. http://dx.doi.org/10.2151/jmsj.2004.895

[3] Brandes, E.A., Ikeda, K., Zhang, G., Schonhuber, M. and Rasmussen, R.M. (2007) A Statistical and Physical Description of Hydrometeor Distributions in Colorado Snowstorms Using a Video Disdrometer. *Journal of Applied Meteorology and Climatology*, **46**, 634-650. http://dx.doi.org/10.1175/JAM2489.1

[4] Hung, G., Bringi, V.N., Cifelli, R., Hudak, R. and Petersen, W.A. (2010) A Methodology to Derive Radar Reflectivity-Liquid Equivalent Snow Rate Relations Using C-Band Radar and a 2D Video Disdrometer. *Journal of Atmospheric and Oceanic Technology*, **27**, 637-651. http://dx.doi.org/10.1175/2009JTECHA1284.1

[5] Hung, G., Bringi, V.N., Moisseev, D., Petersen, W.A., Blivend, L. and Hudake, D. (2014) Use of 2D-Video Disdrometer to Derive Mean Density-Size and Ze-SR Relations: Four Snow Cases from the Light Precipitation Validation Experiment. *Atmospheric Research*, **153**, 34-48. http://dx.doi.org/10.1016/j.atmosres.2014.07.013

[6] Zhang, G., Luchs, S., Ryzhkov, A., Xue, M., Ryzhkova, L. and Cao, Q. (2011) Winter Precipitation Microphysics Characterized by Polarimetric Radar and Video Disdrometer Observations in Central Oklahoma. *Journal of Applied Meteorology and Climatology*, **50**, 1558-1570. http://dx.doi.org/10.1175/2011JAMC2343.1

[7] Nurzynska, K., Kubo, M. and Muramoto, K. (2010) 2D Feature Space for Snow Particle Classification into Snowflake and Graupel. *IEICE Transactions on Information and Systems*, **E93-D**, 3344-3351. http://dx.doi.org/10.1587/transinf.E93.D.3344

[8] Kruger, A. and Krajewski, W.F. (2002) Two-Dimensional Video Disdrometer: A Description. *Journal of Atmospheric and Oceanic Technology*, **19**, 602-617. http://dx.doi.org/10.1175/1520-0426(2002)019<0602:TDVDAD>2.0.CO;2

[9] Grazioli, J., Tuia, D., Monhart, S., Schneebeli, M., Raupach, T. and Berne, A. (2014) Hydrometeor Classification from Two-Dimensional Video Disdrometer Data. *Atmospheric Measurement Techniques*, **7**, 2869-2882. http://dx.doi.org/10.5194/amt-7-2869-2014

[10] Ishimoto, M. (2008) Radar Backscattering Computations for Fractal-Shaped Snowflakes. *Journal of the Meteorological Society of Japan*, **86**, 459-469. http://dx.doi.org/10.2151/jmsj.86.459

[11] Maruyama, K. and Fujiyoshi, Y. (2005) Monte Carlo Simulation of the Formation of Snowflakes. *Journal of the Atmospheric Sciences*, **62**, 1529-1544. http://dx.doi.org/10.1175/JAS3416.1

[12] Tolle, C.R., McJunkin, T.R. and Gorsich, D.J. (2003) Suboptimal Minimum Cluster Volume Cover-Based Method for Measuring Fractal Dimension. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 32-41. http://dx.doi.org/10.1109/TPAMI.2003.1159944

[13] Russ, J.C. (2011) The Image Processing Handbook. 6th Edition, CRC Press, Boca Raton, 604-610. http://www.crcpress.com/product/isbn/9781439840450

[14] Vapnik, V. (1998) Statistical Learning Theory. Wiley, New York.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.