Analysis and Prediction of the Typhoon from an Informatics Perspective

Asanobu KITAMOTO
National Institute of Informatics
2–1–2, Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
kitamoto@nii.ac.jp

Abstract

Analysis and prediction of the typhoon has been intensively studied by a number of meteorologists because of the huge impact of the typhoon to the society. We study the same issue from a different viewpoint — from an informatics perspective. Our goal is to discover relevant knowledge for typhoon analysis and prediction by means of various computational tools that have been developed in the informatics community. Our research takes advantage of the large collection of typhoon data, especially the satellite images of the typhoon, with the application of multimedia data mining methods in the hope of discovering hidden regularities and anomalies in the data collection using data mining algorithms such as principal component analysis, K-means clustering, and self-organizing map. In this paper, we summarize our approaches, achievements and open problems, with the brief introduction of our hand-crafted system, IMET (Image Mining Environment for Typhoon analysis and prediction).

Keywords: Typhoon analysis, Typhoon prediction, Informatics perspective, Typhoon data mining, Image mining environment

> If there is a hurricane, you always see the signs of it in the sky for days ahead, if you are at sea. They do not see it ashore because they do not know what to look for, he thought. — Ernest Hemingway, In *The Old Man and the Sea* (1952)

1 Introduction

If you are a fisherman, the presence of a typhoon, or a hurricane, is of vital importance — an encounter with even a small typhoon may lead to a matter of life and death in the middle of the ocean. Hence a fisherman starts to learn how to find the signs of a typhoon out of, for example, the complex patterns of clouds and their movement in the sky. In other words, a fisherman, or an expert, knows *what to look for* in the sky[1]. On the other hand, a person who lives ashore cannot see those signs even if they actually appear in the sky.

At the moment, computers, or computer programs, are like a person who lives ashore — they do not know yet about *what to look for*, *where to look*, and *how to look* in the sky. This is a significant loss to the society, since the accurate analysis and appropriate forecast of the typhoon has great benefit to the society for the prevention and reduction of natural disasters caused by the typhoon. The goal of our project, *Digital Typhoon project*, is therefore to make computers learn how to see the signs in the complex pattern of typhoon clouds. We, however, do not have to bring computers to the sea in order for them to see the clouds flying in the sky — instead, we can use satellite imagery where you can see the entirety of typhoon cloud patterns from the space.

Problems of the typhoon have been extensively studied in the meteorology community, but we believe that the informatics community can also make a substantial contribution to these problems with the concept of "learning from data." We can take advantage of both sophisticated machine learning technology and powerful computing resources that have made significant advances in the last decade, and these advances have spurred us to discover knowledge from very large databases [1]. We have constructed such very large database of typhoon images, namely *typhoon image collection*, which is the comprehensive image archive of approximately 34,000 typhoon images of consistent quality [2]. This image collection is thoroughly examined by means of various data mining approaches, such as principal component analysis, K-means clustering and self-organizing map, with the aim of discovering regularities and anomalies that may be hidden in the feature space of typhoon cloud patterns. In this paper, we summarize the meteorological background behind those challenges, and some of the methods used for typhoon data mining, and finally introduce our system called IMET (Image Mining Environment for Typhoon analysis and prediction).

[1] Or, only a fisherman who knows *what to look for* in the sky can survive.

2 An Informatics Perspective

2.1 Meteorological Background

Because of the huge impact of the typhoon to the society, the development of typhoon analysis and prediction methods have been one of the primary concerns among meteorologists, and in 1970s, the standard procedure for typhoon analysis called the Dvorak method [3] was established, after nearly 20 years from the first launch of a meteorological satellite in 1960. The basis of this method is the mountains of satellite images we received from many meteorological satellites, and Dvorak established the empirical method by observing numerous satellite images of typhoons. This method, since its inception, has been used in tropical storm analysis centers worldwide.

The Dvorak method is in charge of the interpretation of typhoon cloud patterns on satellite images among the whole procedures of typhoon analysis and prediction. It is essentially a heuristic summarized in the following. Its main components consist of a set of empirical rules that relate various cloud features to a set of parameters representing the intensity of the typhoon, such as central pressure and maximum wind. Those empirical rules are defined for each prototype, incarnated as a sketch drawing that represents a *typical* cloud pattern in a conceptual form. In the analysis stage, forecasters search for similar patterns in the list of sketch drawings and choose the most similar pattern to the real cloud pattern of interest. They then apply empirical rules assigned for the chosen typical pattern, thereby obtaining the intensity estimate of the typhoon at study. In short, the Dvorak method assists human experts for interpreting satellite images and making decisions on the intensity of the typhoon.

The result of interpretation according to the Dvorak method is then fed into a numerical weather prediction (NWP) system, and finally it computes the future evolution of synthetic typhoons that live in the lattice of a three dimensional earth model in a NWP system. This sounds like a totally coherent system, but in fact the Dvorak method has several weak points on which the informatics-based approaches can make an improvement. First, this analysis method is a collection of empirical rules and lacks theoretical background or statistical justification. Second, the potential improvement of the Dvorak method through learning from historical data has largely been unexplored. Most of the valuable satellite data are left unused mainly because of its volumetric challenges to computing and human resources. Third, the Dvorak method relies heavily on the pattern recognition of human experts, and its performance is, for the better or worse, dependent on the capability of human experts' pattern recognition, which is subjective in nature.

The above arguments remind us of a similar framework in the informatics community such as content-based image retrieval and case-based learning, or we may reach more principled understanding of the Dvorak method in the framework of pattern recognition. At the same time, however, we can see the intrinsic difficulty of this procedure from an informatics viewpoint; for example, the direct comparison of *clean* sketch drawings with *noisy* real cloud patterns requires highly *semantic similarity* and intelligent image analysis. Hence it is better to formulate those typhoon problems in a way suitable for computational tools, rather than just simulate the whole procedures in the same way as meteorology or the Dvorak method. This is the motivation we start this research — we challenge typhoon analysis and prediction problems by taking advantage of tools and ideas developed in the informatics community.

2.2 Challenges to Informatics

This is a large-scale real world problem with significant societal impact, and this poses significant challenges to the informatics community in terms of the following research issues:

1. **Spatio-temporal techniques** Observation frequency of every hour generates time series satellite images which are spatio-temporal in nature [2]. Techniques for nearly free-form patterns with complex spatio-temporal dynamics are relatively unexplored areas of research, see for example [4].

2. **Robust techniques** Techniques should be robust enough to deal with every typhoon cloud pattern that could be generated according to the physical laws of the atmosphere. They should also be robust against the complexity of the problem such as computational complexity.

3. **Discovery techniques** The amount of data we receive from satellites is literally explosive because of the recent trends toward more and more sensors with higher and higher spatial, temporal, and bandwidth resolution. This results in satellite data beyond our processing capabilities, hence an important challenge is to develop powerful techniques that digest such large amount of data and uncover hidden information in the dataset.

Solutions to these challenges may lead to new robust spatio-temporal discovery techniques with possible applications in other domains. We do concentrate on this specific domain and build a set of tools effective for this application, but at the same time, we do not make our tools overfit to this application, and try to generalize our tools so that they are applicable to other domains.

2.3 Knowledge-Based Approaches

To the author's knowledge, this is the first attempt from the informatics community toward the comprehensive study of the typhoon. Meteorology in general, however, is not a new application area for the informatics community. In fact, the application of artificial intelligence to the meteorology domain has once flourished in 1980s, when knowledge-based expert systems were in vogue.

[2] Typhoon images could be transformed as volume-like three-dimensional data by estimating the height of clouds.

Table 1: Comparison of our approach with other approaches in terms of models.

Approaches	Models
Meteorology-based approaches	Models of the world (physical processes)
Knowledge-based approaches	Models of the expert (cognitive processes)
Our approach	Models of the data (data generating processes)

One example is a comparative study of artificial intelligence (AI) systems in terms of forecasting severe weather performed during the summer of 1989 [5]. Six systems participated; three traditional expert systems, a hybrid system including a linear model augmented by a small expert system, an analogue-based system, and a system developed using methods from the cognitive science/judgment analysis tradition. According to the authors, however, "*this forecast task turned out to be more difficult than we anticipated; none of the systems produced particularly skillful forecasts.*" Following these unsatisfactory results, recent research found its way to an assistance to experts; for example a system prototype for a tropical cyclone track forecast [6] provides a computer tool for assisting experts by arranging existing meteorological knowledge into an easily understandable form.

2.4 Different Modeling Targets

Now we compare a few relevant approaches in Table 1 in terms of the target of modeling in those approaches. Knowledge-based approaches, as described above, concern the modeling of the cognitive processes behind decision making by the expert. Their results, however, were unsatisfactory probably due to the complexity of both meteorological processes and cognitive processes. On the other hand, meteorology-based approaches concern the modeling of physical processes behind meteorological phenomena, but their scientific approaches are less effective for handling uncertain situations when exact physical models and their simulations are intractable, as is often the case with the typhoon. An example of a meteorology approach is numerical weather prediction, whose success depends on both computing power and the synthetic model of the world that takes as many factors as possible. We point out here two drawbacks in pure meteorology approaches: the incompleteness of typhoon models and the reconstruction of initial conditions.

The first drawback is related to a general attitude that pure meteorology-based approaches cannot deal with situations where solid theoretical models are known just incompletely, because, in meteorology-based approaches, every model should be related to a theoretical foundation that describe the physics of the nature. On the other hand, our informatics approaches generally assume that the true model of the nature is not known, hence we can learn various models from the observed data that approximate the true model of the nature.

Another drawback is related to the reconstruction of the atmosphere. There are two unfortunate situations in the typhoon: ground-truth observations are usually sparse on the ocean, where the typhoon is found most frequently, and the variability of the atmosphere is usually localized around the center of the typhoon. Generally speaking, the reconstruction of atmospheric conditions from sparse observation data is an ill-posed problem and requires some regularity conditions for appropriate reconstruction. However, the smoothness assumption of the atmosphere is usually violated around the center of the typhoon, hence the reconstruction of the atmospheric conditions of the typhoon is a very hard task.

This is an important problem since, without the successful reconstruction of the atmosphere, we cannot simulate the *actual* typhoon found in the atmosphere, even if we have perfect typhoon models. We may be able to apply *analysis-by-synthesis models* to simulate *possible* typhoons, but this does not mean that we can deal with the current typhoon, which is a realization from the space of possible typhoons. In other words, this problem is closely related to the problem of *assimilation* in numerical weather prediction, where the quality of initial conditions affects significantly to the final prediction performance of the system, and the setting of initial conditions require accurate reconstruction of the current atmosphere. Here we conjecture that our informatics approaches may give better initial conditions based on the large collection of historical patterns of typhoon clouds.

2.5 Our Approach : Data mining / KDD

We concern *data themselves*, or the data generating processes (DGP); that is, the modeling of the probabilistic and statistical properties of the observation data. We learn from *data*, not from *humans*, through the modeling of DGP in nature instead of the modeling of cognitive processes in humans. However, our study is not limited to the mere simulation or description of the nature. In fact, our main target is the modeling of DGP that involve both physical factors in generating the data and human factors in observing the data, because both physical factors and human factors serve as prior information for the modeling of DGP.

Through the modeling of DGP, our final goal is to extract relevant information from the large collection of observation data, and derive useful knowledge that extends existing meteorological domain knowledge, or that reveals hidden spatio-temporal regularities and anomalies of the typhoon yet unknown to meteorologists. In short, our approach is based on *data mining*, or *knowledge discovery from databases (KDD)*.

Knowledge discovery from databases concerns knowledge discovery processes applied to databases. KDD deals with ready data, available in all domains of science and in applied domains. Typically, KDD has to deal

Table 2: Relevant research fields for typhoon analysis and prediction.

Image Analysis, Computer Vision	*Image analysis and computer vision is the basic discipline for the extraction of relevant visual features from typhoon images.*
Image Database Systems, Content-Based Image Retrieval	*Management of the large collection of typhoon images requires image database systems. The insertion and deletion of data is normally a trivial task because of the nature of satellite observations, but a support for content-based image retrieval on image features is the key for instance-based learning.*
Pattern Recognition	*Pattern recognition is the basis for treating a pattern as information. We, in particular, rely on statistical pattern recognition whose goal is in representing the probability distribution of the feature space derived from images.*
Artificial Intelligence, Cognitive Science	*Expert systems were unfortunately ineffective as introduced in Section 2.3, but human factors cannot be neglected from our project, since, in any case, we should learn from human experts' pattern recognition — the only successful system in the world on typhoon analysis.*
Meteorology	*Meteorology has established powerful techniques for numerical weather prediction, and quantitative / qualitative knowledge on the physical aspect of the atmosphere. Nevertheless, meteorological methods for the* assimilation *of satellite data into the numerical weather prediction system are still in their infancy.*
Physics, Fluid Dynamics, Chaos	*The dynamics of the atmosphere can be well described by a set of physical laws, and fluid dynamics play an important role in this framework. However, we should always pay attention to the concept of* chaos*, which is an indispensable concept when predictability is concerned.*

with inconclusive data, noisy data, and sparse data. Data mining indicates the application of low level data mining methods under human control, where data mining methods are algorithms designed to analyze data, or to extract patterns in specific categories [1]. This paper summarizes our current *exploratory* effort in search of effective data mining / KDD methods and algorithms for this particular application, the typhoon image collection.

Other recent informatics-based approaches on the typhoon (hurricane) include [7, 8] and several others. Those references discuss the application of active contour models, optical flow, neural networks and fuzzy logic to the analysis of the typhoon. The most significant difference between our research and those researches, however, is that our research is based on the consistent and comprehensive large data collection of typhoon satellite images. In contrast, their standpoint is the application of particular informatics methods to the meteorology domain based on miniature datasets; hence their impact on meteorology seems to be limited. Other research fields relevant to typhoon problems are summarized in Table 2.

2.6 Discovery of the Signs

Normally the numerical weather prediction system produces relatively good prediction performance because simple extrapolation is good enough for typhoons that strengthen or weaken slowly. In some situation, however, this scheme cannot provide good forecast because of sudden changes of the typhoon. To name a few:

1. Sudden and rapid intensification,
2. Irregular or random movement,
3. Typhoon formation or cyclogenesis,

may be the representative cases of sudden changes of the typhoon. In fact, meteorologists have neither solid understanding nor good prediction performance for these phenomena. This situation, in turn, indicates a possibility that informatics approaches may be able to make a contribution to these problems with a different viewpoint.

Thus we are interested in *difficult* problems that have not been answered by the meteorology community. Our goals are not in improving overall prediction performance but in discovering the *signs* in cloud patterns that predate the occurrence of such rare phenomena. Our goal is to see those signs in the spatio-temporal features of typhoon image sequences.

3 Typhoon Image Collection

At the moment, the typhoon image collection archives more than 34,000 *well-framed* typhoon images as summarized in Table 3. Here the term *well-framed* means: (1) The center of the typhoon is always registered with the center of the image. (2) The image captures most of the typhoon cloud system with the minimal effect from distortion in shape and size. The typhoon center is determined from the *best track* dataset that will be introduced later. Thus the data collection as a whole provides a medium-sized, richly-variational, and carefully-preprocessed scientific data collection with real applications. Hence it can be used as an interesting large-scale testbed for spatio-temporal data mining.

Our collection is comparable in size to similar hurricane archives under development at NESDIS/CIRA (National Environmental Satellite Data and Information Service / Cooperative Institute for Research in the Atmosphere) in USA [9]. The collection consists of 40,000+

Table 3: The current status of the typhoon image collection.

Basin	Northern Hemisphere	Southern Hemisphere
Best Track		
Name of agency	Japan Meteorology Agency (JMA)	Australian Bureau of Meteorology (BOM)
Latitudinal Domain	$0^{\circ}N \sim$	$\sim 0^{\circ}S$
Longitudinal Domain	$100^{\circ}E \sim 180^{\circ}E$	$90^{\circ}E \sim 170^{\circ}E$
Typhoon Image Collection		
Typhoon seasons	6 Seasons (1995–2000)	5 Seasons (1995–2000)
Number of sequences	136	62
Number of images	24,500	9,400
Images per sequence	53 ~ 433	25 ~ 480
Observation frequency	1 hour	1 hour

images for tropical cyclones in the Atlantic and the eastern Pacific for the period 1996-2001. Their background is in meteorology, but they point out that, even in the meteorology community, there have been few quantitative applications of satellite imagery for investigations of intensity, structure, and motion of the hurricane, particularly with large data samples. Thus our research shares some motivations with their research, but they are concerned mainly with the analysis of the hurricane archive by traditional statistical analyses.

4 Typhoon Data Mining

4.1 Categorization of Data Mining

Typhoon data mining we deal with in this paper is a data mining for a scientific domain on a meteorological application in the form of image / multimedia data with spatio-temporal properties. Hence many types of data mining algorithms can be applied to this data collection, and we may need an extensive and comprehensive study to determine which algorithms work best for this particular application. To review various data mining algorithms, we classify them into three categories: spatial data mining, temporal data mining, and spatio-temporal data mining.

Spatial data mining deals with the two-dimensional distribution of typhoon cloud patterns, but note here that a feature space for two dimensional spatial patterns has, in general, much higher dimensions than two. Temporal data mining, on the other hand, focuses on the temporal dynamics of typhoon cloud patterns and involves the modeling of the life cycle of the typhoon. Relatively speaking, spatial data mining is more concerned with typhoon analysis, while temporal data mining is more concerned with typhoon prediction. Spatio-temporal data mining integrates both types of data mining, and therefore should be most powerful, but we are yet to develop or test algorithms of this category because of the complexity of the data collection and algorithms.

Toward the mathematical models of spatial patterns of the typhoon, we purse two approaches, namely component-based and shape-based representation. First, in component-based representation, we investigate an approach that represents typhoon cloud patterns with the weighted combination of basic *components*. Here a component represents the distribution of clouds which is characteristic for the dataset at study. This approach does not require the segmentation of an image, hence this is robust but still powerful. We begin with PCA (principal component analysis), which is an orthodox mathematical method for the efficient reduction of dimensionality while retaining maximum variability in the dataset. The applications of PCA to image datasets include face recognition [10] and remote sensing images [11], and in the context of meteorology, PCA is often used with the name EOF (empirical orthogonal function) [12].

On the other hand, in shape-based approaches, we explicitly represent cloud patterns with mathematical shape models. An example of this approach is a shape decomposition method for representing typhoon cloud patterns with a set of ellipses [15]. Here an ellipse is used as a basic component because an ellipse and a spiral corresponds to meteorologically meaningful parts of the typhoon, cloud clusters and spiral rainbands, respectively. Thus the explicit representation of those elements leads to effective image features for content-based image retrieval. More principled approaches to shape-based representation include the probabilistic model of shape, see [4]. in data mining.

4.2 Dimensionality Reduction

The application of PCA has two purposes: namely the extraction of components and the reduction of dimensionality. The first purpose corresponds to extracting eigenvectors that represent maximum variability contained in the dataset, and these eigenvectors are often called "eigen-X" depending on the application. In our application, an eigenvector may be called an *eigen-typhoon*. Figure 1 represents eigen-typhoons for the northern and the southern hemisphere, with the average typhoon and the variance typhoon. The first principal eigen-typhoon represents the difference of cloud fraction between the northern and the southern part of the image, or the latitudinal structure of the typhoon. Eigen-typhoons with smaller eigenvalues represent spiral components that look like rainbands. Thus these images represent the typical distributions of typhoon cloud patterns. Next the lower panels of Figure 1 represent the cumulative proportion of eigen-

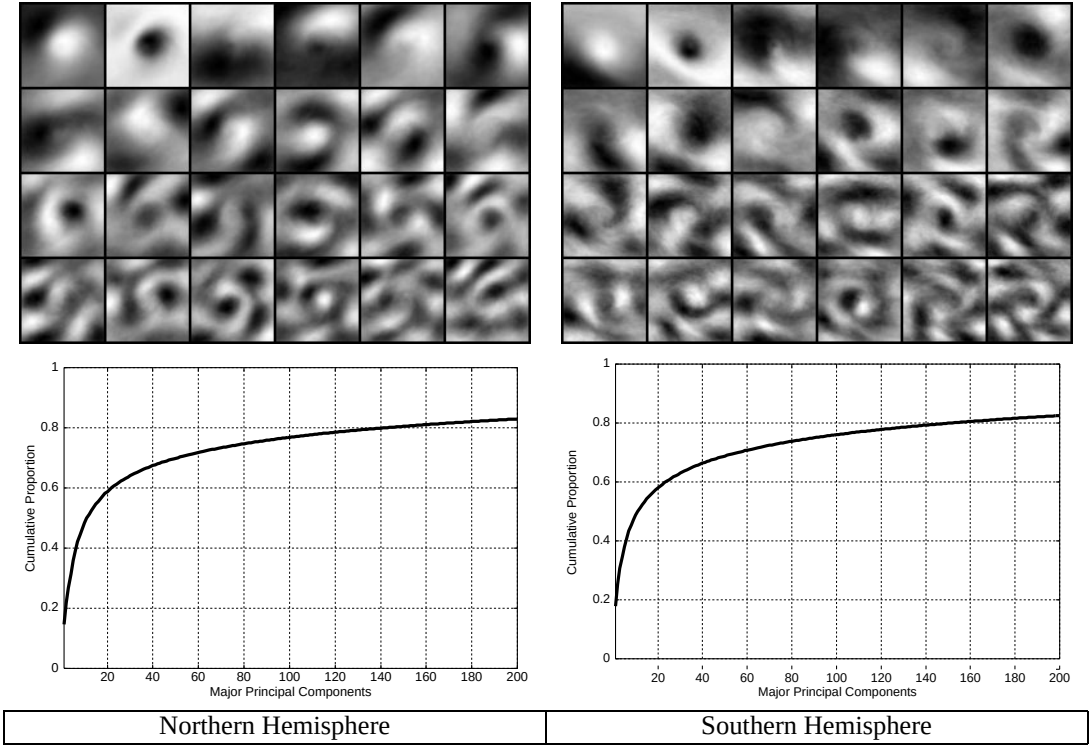


Figure 1: The eigenvectors of typhoon cloud patterns, or *eigen-typhoons*, for the northern and the southern hemisphere. From upper-left corner: the mean image, the variance image, and eigenpictures from the 1st to the 22th. The cumulative proportion is also illustrated in the bottom panels.

values. About 75% of the variability contained in the dataset can be represented with only an 83 dimensional vector for the northern hemisphere out of a $64 \times 64 = 4096$ dimensional vector. The same threshold is 91 for the southern hemisphere.

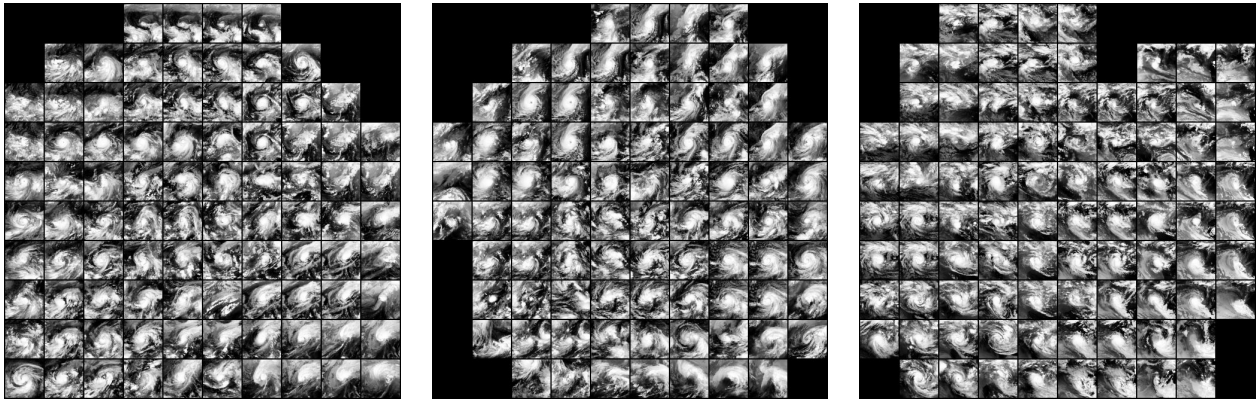
4.3 Projection into Low Dimensions

The intuitive visualization of a high dimensional feature space can be obtained by *projection*. Linear projection methods simply transform a data point in a high dimensional feature space into a lower dimensional latent space with a linear projection matrix, and nonlinear projection methods, including clustering approaches as discussed in the next subsections, may provide more compact representation. We first visualize the feature space with principal components that we derived in the previous subsection. Figure 2 illustrates the projection of data points into two dimensional spaces, whose axes represent the first and the second principal components. In these spaces, grid points are chosen with uniform intervals along each axis, and those points are then projected back to the original high-dimensional space using a Moore-Penrose generalized inverse matrix. Then the nearest image to each grid point is visualized on the two dimensional array of grid points as in Figure 2. These figures suggest the implication of principal components: the first principal component represents the latitudinal structure, while the second principal component is more closely related to the orientation of the major cloud region. We can test other linear projection matrices, and in fact, even a randomly chosen linear projection matrix results in a two dimensional visualization that shows the spatial distribution of cloud patterns. Hence, a potential research area is to find an interesting projection matrix, or *projection pursuit*.

4.4 Discovering Typical Patterns

Clustering procedures aim at yielding a data description in terms of clusters or groups of data points that possess strong internal similarities [13]. For the typhoon image collection, we expect that clustering procedures may produce the intuitive summarization of typhoon cloud patterns that can be used as the *catalog* of typhoon images. If we can find a set of clusters that represent typical patterns of the typhoon, we can categorize complex cloud patterns into several representative patterns, thereby characterize them with a set of basic patterns. The Dvorak method, introduced in Section 2.1, did this task manually and selected typical cloud patterns embodied from the long experience of analysts. In contrast, we do this task automatically.

The basic non-hierarchical clustering procedure is the K-means clustering, and the result of clustering is illustrated in Figure 3. In this experiment, the number of clusters is fixed to 100, and clusters obtained through experiments are shown in no particular order. Those images can be considered as representative patterns, and many types of shape are visualized together on a two dimensional space. From another viewpoint, this is the nonlinear projection of the high dimensional feature space. Hence it is a concise visualization, but it is still not an



(a) PCA 1-2 (N.H.) (b) PCA 3-4 (N.H.) (c) PCA 1-2 (S.H.)

Figure 2: Projections of the feature space into two dimensional space spanned by leading principal components. Horizontal and vertical axis represents leading principal components as specified in the caption below the panels, and N.H. and S.H. represents the northern hemisphere and the southern hemisphere, respectively.

intuitive visualization for humans due to the lack of ordering between clusters.

In this sense, better representation in terms of the ordering of clusters can be obtained through the SOM clustering. It summarizes high dimensional data vectors with a set of reference vectors having a topological organization on a (usually) two-dimensional lattice. The detail of the algorithm is found in many publications [14], so we only describe some settings we use for the basic SOM. The array of neurons is configured on a square lattice of either the lattice or the torus topology. Topological neighborhood is defined in reference to chess-board distance on a square lattice, and learning rate factor is proportional to the inverse of the number of steps with some minimum limit. Reference vectors are randomly initialized.

Figure 4 (a) and (b) illustrate results of the SOM clustering. They give an improved visualization with apparent spatial ordering. These clustering methods can thus visualize the high dimensional feature space of typhoon cloud patterns in a "birds-eye-view" representation, which is effective for understanding the overall distribution at a glance. Thus the ordering of typhoon cloud patterns attained by the SOM gives a unique insight into the nature of typhoon cloud patterns.

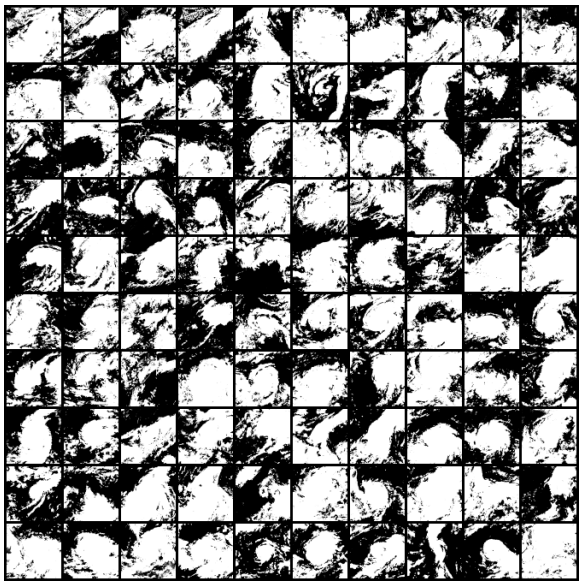


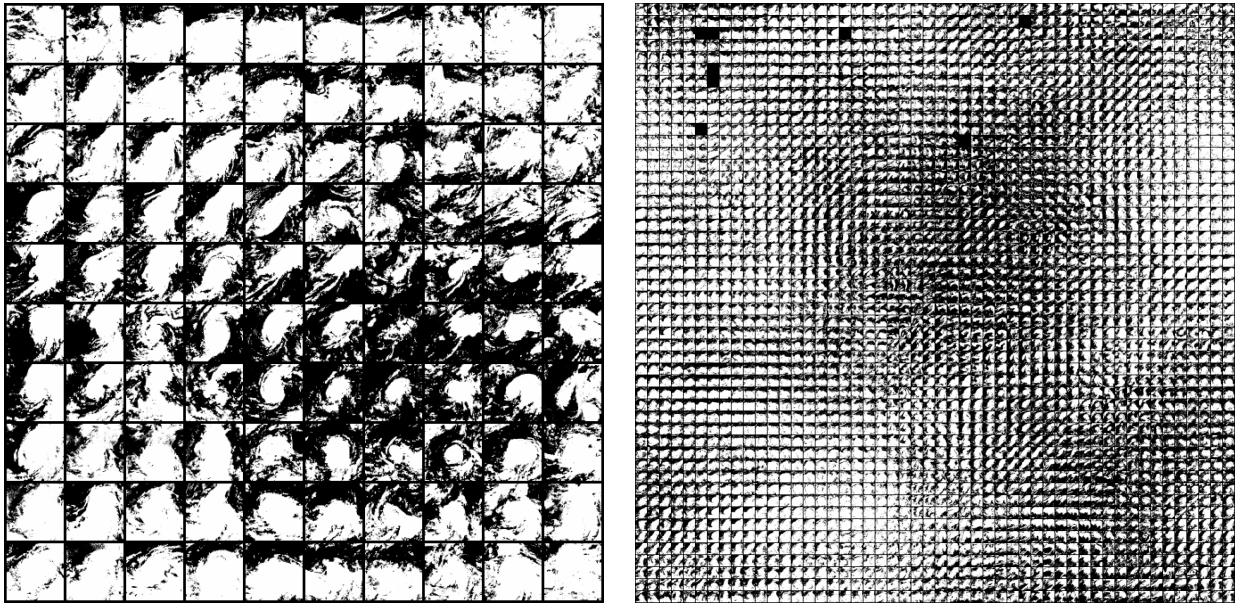
Figure 3: K-means clustering of typhoon cloud patterns. Clusters are visualized in no particular order.

4.5 Temporal Typhoon Data Mining

We next consider the modeling of temporal aspect of typhoon cloud patterns. The most intuitive approach may be to learn from history or to apply case-based learning. Here, case-based learning represents knowledge in terms of specific cases or experiences and relies on flexible matching methods to retrieve these cases and apply them to new situations [16]. Based on this concept, we can imagine a typhoon prediction scenario by case-based learning:

1. We first create the *well-framed* image of the typhoon at study, and use this image as the query example to the typhoon image database.
2. The typhoon image database performs a similarity-based image retrieval and returns a list of similar images found in the past.
3. We then refer to the evolution of past similar typhoon sequences, and predict the typhoon at study based on some inference on the *ensemble* of past similar cases.

This scenario might seem to be reasonable, but, in reality, this scenario is challenging from both a practical and a theoretical viewpoint. First, this scenario is practically challenging because similarity between cloud patterns involves *semantic similarity* such as the presence of the eye. Second, this scenario is theoretically challenging because typhoon prediction, or the prediction of atmospheric events in general, cannot escape from the fundamental issues of predictability.



(a) SOM on 10×10 nodes (b) SOM on 50×50 nodes

Figure 4: Clustering of typhoon cloud patterns using the SOM clustering.

The issue of predictability, or short-term predictability and long-term unpredictability, focuses on the nonlinear dynamical processes of the atmosphere. This issue in the context of case-based prediction of atmospheric situation was in fact studied more than 30 years ago by one meteorologist, who is well known for his discovery of *chaos*, whose deep insight finally uncovered the unpredictable nature of the atmosphere. In his pioneering work [17, 18], he tried to find similar weather situations (analogues) in terms of the pressure pattern of the upper troposphere obtained from historical weather data in the hope of utilizing historical data for the prediction of the current weather. However, the result was disappointing. He found that similar weather situations rapidly lead to dissimilar situations and he insisted that there were indeed no truly good analogues. His opinion is that in practice case-based prediction might be expected to fail. Afterwards, similar findings have been repeatedly reported in the meteorology community. This pessimistic outlook on case-based prediction could be more optimistic with powerful typhoon models and associated similarity metric that focuses on truly relevant image features.

5 IMET Overview

5.1 System Architecture

Our arguments so far indicate that typhoon problems are highly complex problems, and our data mining approaches are yet to make a substantial contribution to these problems. However, more immediate contributions can be made through the construction of an intelligent information system for the typhoon image collection, where the informatics-based approach can naturally play an important role. We therefore build the system that we call IMET (Image Mining Environment for Typhoon analysis and prediction), which is designed for the intelligent and efficient searching and browsing of the typhoon image collection.

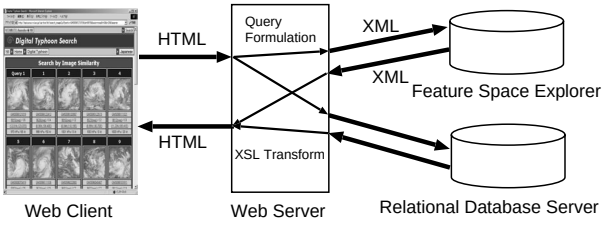


Figure 5: Distributed system architecture of IMET.

Figure 5 illustrates the overall architecture of IMET. The system consists of three main components: Web browser clients as the user interface, the Web servers which may act as hierarchical meta servers, and backend database servers where typhoon data are actually archived. Moreover, those components may be distributed over the network to allow distributed database systems, which is often the case with satellite data archives. At the moment, we have two types of backend database servers, namely relational database management systems (RDBMS) and our hand-crafted image search engine called FSE (Feature Space Explorer). For this architecture, we need to prepare two types of languages:

1. **Query language** Describe a query from clients or Web servers to backend database servers.

2. **Definition language** Describe the contents of data archived in backend database servers.

The creation of such languages has been an active area of research, and there have been numerous proposals for new languages that are designed for specific purposes. A

Query specification

1. A single query example is chosen randomly from Typhoon 9903, and images that belong to this typhoon sequence is filtered out from subsequent tasks.
2. Images in the database are grouped by the name of typhoon sequences. Distance to the query example is calculated for each image, and images in each group are then sorted by distance in ascending order.
3. Fetch at most 2 similar images from each group. Those images are collected into the parent group, and again sorted by distance in ascending order. Finally top 5 images are fetched from the parent group, resulted in 5 most similar images in which at maximum 2 images are fetched from one typhoon sequence.
4. Return the list of similar images with the name of the typhoon sequence, the name of the image, and distance between the query example and each image, and the query example of this task.

XML encoding of a query

```
<?xml version="1.0" encoding="UTF-8"?>
<envelope>
 <header>
  <server port="59300">localhost</server>
  <session user="kitamoto" id="1">
   <transaction>1</transaction>
  </session>
 </header>
 <body>
  <query>
   <task>
    <example type="single">
     <constant select="folder">9903</constant>
     <dynamic select="name">@random</dynamic>
    </example>
   </task>
   <where>
    <filter select="folder" type="equals" not="1">@example</filter>
   </where>
   <sort-by order="ascending">value</sort-by>
   <fetch>
    <from>0</from>
    <size>5</size>
   </fetch>
   <return>
    <select>folder</select>
    <select>name</select>
    <select>value</select>
    <select>example</select>
   </return>
   <group-by>
    <select>folder</select>
   </group-by>
   <for-each>
    <task>
     <let variable="value">
      <function target="example">distance</function>
      <metric type="euclid" option="squared">
       <min>0</min>
       <max>30</max>
      </metric>
     </let>
    </task>
    <sort-by order="ascending">value</sort-by>
    <fetch>
     <from>0</from>
     <size>2</size>
    </fetch>
   </for-each>
  </query>
 </body>
</envelope>
```

XML encoding of a result

```
<?xml version="1.0" encoding="UTF-8"?>
<envelope>
 <header>
  <session user="kitamoto" id="1">
   <transaction>1</transaction>
   <matching>24300</matching>
   <elapsed>0.000000e+00</elapsed>
  </session>
 </header>
 <body>
  <example>
   <folder>9903</folder>
   <name>GMS5990601l3</name>
  </example>
  <list number="5">
   <item order="0" id="0">
    <folder>9902</folder>
    <name>GMS599042809</name>
    <value>1.962298e+00</value>
   </item>
   <item order="1" id="1">
    <folder>9514</folder>
    <name>GMS5950919O8</name>
    <value>3.230482e+00</value>
   </item>
   <item order="2" id="2">
    <folder>9915</folder>
    <name>GMS5990916O6</name>
    <value>3.372034e+00</value>
   </item>
   <item order="3" id="3">
    <folder>9509</folder>
    <name>GMS595082415</name>
    <value>4.487362e+00</value>
   </item>
   <item order="4" id="4">
    <folder>0003</folder>
    <name>GMS500070219</name>
    <value>5.203874e+00</value>
   </item>
  </list>
 </body>
</envelope>
```

Figure 6: A query specification, its XML encoding, and a result in XML encoding.

representative example is SQL (Structured Query Language), which is the standard query language for RDBMS and its variants. In terms of multimedia applications, MRML [19] is proposed as a query language with the open communication protocol for CBIR in a XML-based markup language, and MPEG-7 is also proposed as a definition language with a large vocabulary for the description of multimedia contents. Nevertheless, it seems that we are yet to reach the standard language for multimedia applications, and this is the motivation that we develop our hand-crafted prototype languages for the query and the definition. The advantage of having such hand-crafted languages is in rapid prototyping of new tasks required for typhoon data mining. Our intention is not in developing full-fledged languages with rigorous theoretical foundations. Instead, the design goal of our languages is to create a handy yet useful languages with maximally orthogonalized operators whose combination describe various actions needed in IMET.

Figure 6 describes a query specification in the upper part of the table, and its XML encoding in the left side, and a result in the right side. As Figure 6 shows, our query language relies on XML for the syntax of the language, and also relies on XQuery (Query Language for XML), or its full XML-encoded XQueryX (XML Syntax for XQuery), and other XML-related standards for the semantics of the language. The query in Figure 6 corresponds to a little complex query-by-example similarity-based retrieval, in which the grouping of data by typhoon name is contained in a way similar to a sub-query. Once a query is formulated, we submit this XML message to the image search engine, FSE, and the engine then returns the result of the specified task encoded also in XML.

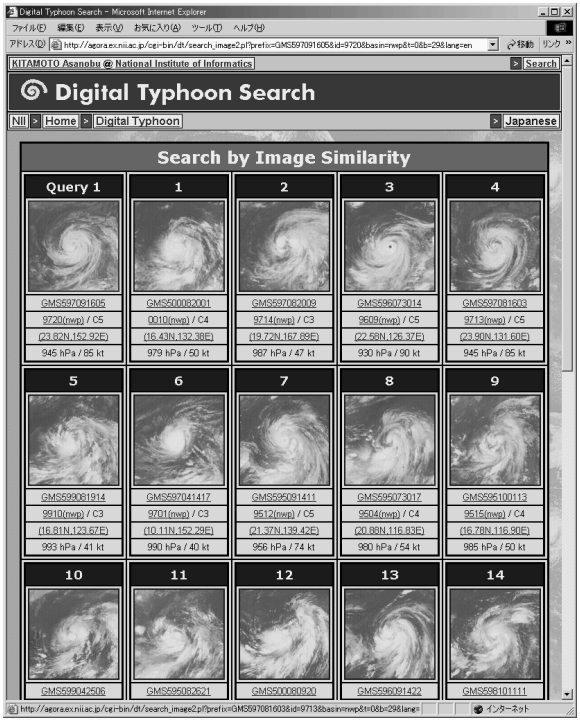


Figure 7: Nearest neighbor searches in IMET.

5.2 Results

IMET supports traditional searching functionality based on metadata, such as *search-by-name*, *search-by-date*, and *search-by-geography*. In addition, IMET provides K-NN (nearest neighbor) similarity-based retrieval with Euclidean metric to search for similar patterns in the past as shown in Figure 7. IMET also provides the comparison of multiple typhoon sequences mainly for prediction purposes. Our system IMET is now open to the public and can be accessed at the Web site *Digital Typhoon* with the URL http://www.digital-typhoon.org/ so that interested readers can explore the typhoon image collection and the IMET system.

6 Conclusion

We introduced our research on typhoon analysis and prediction from an informatics perspective. Our principle is data mining and case-based learning, and we showed several approaches that we applied to the large collection of typhoon images. The ideas behind those approaches are to study the regularities and anomalies of typhoon cloud patterns in the feature space through exploiting our principle — *learning from data*. The results in this paper are in the preliminary stages in the sense that we are yet to derive some definitive knowledge useful for typhoon analysis and prediction. However, we are accumulating our experience with typhoon images through our developed system IMET that we briefly introduced in the last part of this paper. We are planning to implement more powerful learning technology, and more effective search engine with an efficient query language.

Acknowledgment

We express our deep appreciation to Prof. M. Kitsuregawa and Prof. Y. Yasuoka of Institute of Industrial Science, University of Tokyo, for granting access to the huge data archives of GMS-5 satellite images. This work was in part supported by the Ministry of Education, Culture, Sports, Science and Technology, Grants 12780300, and by Telecommunications Advancement Organization of Japan (TAO).

References

[1] U.M. Fayyad, G.Piatetsky-Shapiro, P. Smyth, and R. Uthurusammy, editors. *Advances in Knowledge Discovery and Data Mining*. The MIT Press, 1996.

[2] A. Kitamoto. Data mining for typhoon image collection. In *The 2nd International Workshop on Multimedia Data Mining*, pages 68–77, 2001.

[3] V.F. Dvorak. Tropical cyclone intensity analysis using satellite data. *NOAA Technical Report NESDIS*, 11:1–47, 1984.

[4] A. Blake and M. Isard. *Active Contours*. Springer, 1998.

[5] W.R. Moninger, J. Bullas, B. de Lorenzis, E. Ellison, J. Flueck, J.C. McLeod, C. Lusk, P.D. Lampru, R.S. Phillips, W.F. Robers, R. Shaw, T.R. Stewart, J. Weaver, K.C. Young, and S.M. Zubrick. Shootout-89, a comparative evaluation of knowledge-based systems that forecast severe weather. *Bulletin American Meteorological Society*, 72(9):1339–1354, 1991.

[6] L.E. Carr, R.L. Elsberry, and J.E. Peak. Beta test of the systematic approach expert system prototype as a tropical cyclone track forecasting aid. *Weather and Forecasting*, 16:355–368, 2001.

[7] L. Zhou, C. Kambhamettu, and D.B. Goldgof. Extracting nonrigid motion and 3D structure of hurricanes from satellite image sequences without correspondences. In *Proc. of Conference on Computer Vision and Pattern Recognition*. IEEE, 1999.

[8] R.S.T. Lee and J.N.K. Liu. An automatic satellite interpretation of tropical cyclone patterns using elastic graph dynamic link model. *Pattern Recognition and Artificial Intelligence*, 13(8):1251–1270, 1999.

[9] R.M. Zehr. Tropical cyclone research using large infrared image datasets. In *24th Conference on Hurricanes and Tropical Meteorology*, pages 486–487. American Meteorological Society, 2000.

[10] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–86, 1991.

[11] U.M. Fayyad, P. Smyth, N. Weir, and S. Djorgovski. Automated analysis and exploration of image databases: Results, progress, and challenges. *Journal of Intelligent Information Systems*, 4:7–25, 1995.

[12] D.S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 1995.

[13] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[14] T. Kohonen. *Self-Organizing Maps*. Springer, second edition, 1997.

[15] A. Kitamoto. The development of typhoon image database with content-based search. In *Proceedings of the 1st International Symposium on Advanced Informatics*, pages 163–170, 2000.

[16] P. Langley. *Elements of Machine Learning*. Morgan Kaufmann Publishers, 1996.

[17] E.N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26:636–646, 1969.

[18] E.N. Lorenz. Three approaches to atmospheric predictability. *Bulletin American Meteorological Society*, 50(5):345–349, 1969.

[19] W. Müller, Z. Pecenovic, A.P. de Vries, D. Squire, H. Müller, and T. Pun. MRML: Towards an extensible standard for multimedia querying and benchmarking. Technical report, Computing Science Center, University of Geneva, 2000.