行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　適用於科學資料的資料管理與分析系統之研究　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：¬個別型計畫　　□整合型計畫
計畫編號：NSC　89-2213-E009-176
執行期間：89 年　8　月　1　日至 90　年　7　月 31　　日

計畫主持人：梁　婷
共同主持人：

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學資訊科學系

中　華　民　國　　90　年　10　月　22　日

行政院國家科學委員會專題研究計畫成果報告

適用於科學資料的資料管理與分析系統之研究

Study of a data management and analysis system for scientific database

計畫編號：NSC　**89-2213-E009-176**
執行期限：89 **年** 8 **月** 1 **日至** 90 **年** 7 **月** 31 **日**

主持人：梁　婷　　國立交通大學資訊科學系
計畫參與人員：洪炎東、時繼弘、張譽騰 國立交通大學資訊科學系

一、中文摘要

隨著資料庫技術的發展，各式各樣的數據大量的產生。如何進一步將數據轉換成有用的信息是有非常迫切的需要。在本計劃中我們針對大量計算流體力學數據提出並製作一個渦流資料管理與分析系統。此系統將包含四個主要模組設計：渦流特徵萃取、關鍵影像萃取、查詢機制和使用者介面。我們希望本計劃的執行，不僅在知識探勘技術上有進一步探討與應用，也能經由系統的實作提供流力研究者一個好的分析與管理工具。

關鍵詞：資料管理、分析、特徵萃取、關鍵影像、查詢機制、使用者介面

Abstract

With the advent of database technology, various communities have accumulated increasingly large amount of data. Data from various activities need further analysis to transform it into useful information. Hence a data management and analysis system purposely designed for flow data is investigated and implemented in this project. The system contains four main modules, feature extraction, key frame extraction, search engine and user interface. The implementation of this project will benefit both the information scientist in the context of knowledge discovery and at the same time provide an efficient flow data management and analysis tool for researchers in the computational fluid dynamics community.

Keywords: data management, analysis, feature extraction, key frame extraction, indexing mechanism, user interface

二、緣由與目的

Due to the advance in technology, the capabilities of collecting, generating and storing data grow much faster than our abilities to analyze, summarize and extract useful knowledge from them. Although database technology has provided us with the basic tools for efficient storage and lookup for large data sets, it still remains a challenge to analyze and manage large bodies of data effectively [3, 4, 12].

Contrast to high diversity of the information and knowledge of corpus in natural languages [10] most scientific databases are collected with some specific purpose so that the information and the possible knowledge are comparatively restricted and predictable in some aspects. Moreover, the data structure in a scientific database is relatively simpler and they are usually highly correlated. For example, the flow data obtained by researchers of computational fluid dynamics are time varying velocity field composed of three real numbers in each point in space and hence, their temporal and spatial variations are clearly correlated [13]. Therefore, it is possible to design a data management system of large scientific database for efficient information retrieval [12]. Such an idea was initiated in a visit to Vortical Flow Research Laboratory in MIT during 1998 to 1999. Obviously, to attain this goal, collaboration

with experts in computational fluid dynamics is absolutely essential.

In computational fluid dynamics, the Navier-Stokes equation is solved numerically under different physical situation. The outcomes of the computation are the velocity and pressure fields. For a computational domain consists of 100 points in each dimension, in two-dimension there will be 30,000 double precision real numbers (20,0000 for the velocity field and 10,000 for the pressure field) generated at each time step. If a double precision number occupies 16 bytes, the size of a flow field data file will be 500 Kilobyte. Typically, in one single run, calculation of more than 10^6 time steps are needed and 10^{12} byte of data will be generated. For three-dimensional problems, the amount of data will be of order 10^{18} bytes! Thus, it is impossible to record and analyze every calculated flow field for each time step. Usually, only flow field values at some chosen spatial points and flow fields at some chosen time are recorded. The rest of the flow data, which are obtained by more than 99.99% of the computation, are discarded without analysis. One of the reasons for such waste was because massive data storage device was not available in the past.

Today, with the advance in storage technology, disk arrays of 10^{12} byte become affordable and it is possible to store every bit of the flow field data. However, it remains a difficult task to manage and analyze such huge amount of data. Hence, it will be important to develop a flow data management and analysis system that can automatically identify features and extract relevant flow fields for the researcher. In this project, we address the four main issues, namely, features extraction, key frame extraction, frame indexing, and user interface, of a flow field data management and analysis system.

In feature extraction, the features such as flow separation, vortex center, vortex size, stagnation points… etc. are identified from sequences of flow field images by using the techniques of image processing [7]. As to the key frame extraction, several methods have been proposed for various applications in recent literatures [2, 5, 6, 8, 15]. These approaches are based upon detection of disrupted changes to segment a continuous video stream into shots. For example, a difference measurement to evaluate the difference for every pair of adjacent frames can be used to detect shot boundary when the difference is greater than a predefined threshold value. Then either the first frame or the last frame of a shot can be treated as key frames regardless a shot's complexity. More sophisticated approach was proposed on the basis of shot activity [2]. The drawback with this approach is high computational expense. Other approaches based on visual content are proposed in [15] in which multiple visual criteria are considered, for example, color features and motion. Also the extraction of key frames can be done as the extraction of cluster centroids by an unsupervised clustering [15]. In this project, a content-based key frame extraction is used. A difference measurement function is used as shot boundary detection and the number of key frames is proportional to shot complexity.

To support fast retrieval from huge volumns of flow data an effective frame indexing mechanism is required [9, 11]. Usually the indexing scheme implemented in a practical video system is purposely tailored to a certain types of users' inquires. In this project a two-level index structure is constructed to record the feature contents for each flow frame. On the other hand the friendly user interface is constructed to support menu-based and point-click-and-drag visual iconic input queries which are useful for fluid dynamics researchers.

In the end, the goal of this project is to explore possible new and better ways to manage large flow data sets from computational fluid dynamics. On the other hand, this project will benefit both the information scientist in the context of knowledge discovery and at the same time help develop a good data management system for the fluid dynamist to better deal with the flow data

三、結果與討論

The proposed system was implemented in Windows 2000 and CPU P4 (Intel). An image development language IDL (interactive data language) from Research System Inc. is used for displaying and analyzing flow field data. Figure 1 is the brief flowchart of the overall system. The flow field data were obtained from our collaborator Professor Robert Hwang from Naval Structure and Ocean Engineering Department, National Taiwan University.

In a typical flow field, there are many features such as vortexes, vortex shedding, instabilities, turbulence structure, … etc. In this project, the properties of vortex such as position, strength, …etc. are investigated. Since it is easier to work with the stream function [13] instead of the velocity field directly, the preprocessor will first transform the velocity field data into stream values and generate corresponding stream line and velocity field mpeg files for user's browsing. The feature extractor successfully identify significant vortex features by using image processing techniques which the domain knowledge is also encoded. Details of techniques description can be referred in [14]. Currently the features include vortex center, vortex size, stagnation points, separations, …etc.

The extracted features are employed to construct the key frame extractor. Since every frame in a sequence of images corresponds to a specific time instance and is characterized by a specific feature vector, the feature vectors of all images in the sequence form a trajectory which expresses the temporal variation of the image feature vectors of the sequence of images. Therefore, shot boundaries can be identified as those between local minimum and local maximum of feature differences. However those adjacent shots whose feature differences are less than the difference mean will be merged to be one shot. Those shots whose differences larger than the means are treated as shots and the number of key frames per shot is proportional to the difference complexity. In this project there generated 43 shots and 294 key frames for 2000 data set and 188 shots and 1941 key frames for 13000 data set. Table 1 shows the preprocessing time for two data sets.

To support fast search for image frames, the extracted features are used to construct inversion. A two-level shot-to-frame tree structure will be designed to represent the contextual relationship among shots and frames and a summarization record generated by built-in summary function will be assigned to each shot. The interface support menu-based and point-click-and-drag visual iconic input queries as well as Boolean-based queries. A message bar will guide user inquiry by prompting system response description in natural language. Figures 2 to 7 are parts of system demostrations.

四、成果自評

Throughout the implementation of the project, the vortex information system is well constructed. It provides users a very friendly interface for inquiry and management of flow field data. It also provides analysis and browsing tools, making it helpful for user in knowledge discovery. Further improvement can be directed on the identification of principal features used in key frame extraction algorithms. We appreciate financial supports from National Science Concil to the extension of this project on advanced analysis tools development on global feature extraction and clustering methods.

五、參考文獻

[1] G. Ahanger and T. D. Little, (1996) "A survey of technologies for parsing and indexing digital video," Journal of Visual Communication and Image Representation, Vol. 7, No. 1, pp. 28-43.

[2] Y. S. Avrithis, A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, (1999) "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases," Computer Version and Image Understanding, Vol. 75, Nos. 1/2, July/August, pp. 3-24.

[3] C. Bettini, X. S. Wang, S. Jajodia, J. Lin, (1998) "Discovering frequent event patterns with multiple granularities in time

sequence," IEEE Transactions on Data Knowledge and Data Engineering, Vol. 10, No. 2, 1998, pp222-237.
[4] Steven A. Chien and Helen B. Mortensen, (1996) "Automating Image Processing for Scientific Data Analysis of a Large Image Database," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 8, August, pp. 854-859.
[5] N. D. Doulamis, A. D. Doulamis, Y. Avvrithis, and S. D. Kollias, (1999) "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases," Multimedia Signal Processing, 1999, pp. 141-146.
[6] E. Kang, S. Kim, and J. Choi, (1999) "Video Retrieval Based on Scene Change Detection in Compressed Streams," IEEE Transactions on Consumer Electronics, Vol. 45, No. 3, August, pp. 932-935.
[7] A. K. Jain and R. C. Dubes, (1988), Algorithms for clustering data, Prentice-Hall, 1988.
[8] R. L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic, and E. Persoon, (1996) "Visual Search in a SMASH System," Proceedings of Image Processing, Vol. 3, pp. 671-674.
[9] B. I. Lee, Y. I. Chang, and W. P. Yang, (1997) "An efficient conflict-resolution approach to support read/write operations in video server," International Journal of Software Engineering and Knowledge Engineering, Vol. 7, No. 3, pp.1-29.
[10] Tyne Liang and T. C. Wu, (2001) "A linguistic approach to Chinese query model," Int. Journal of Information and Science Engineering, Vol. 17, 95-111.
[11] S. W. Smoliar and H. Zhang, (1994) "Content-based video indexing and retrieval," IEEE Multimedia, pp. 62-72.
[12] Paul Stolorz, and Peter Cheeseman, (1998) "Onboard Science Data Analysis: Applying Data Mining to Science-Directed Autonomy," IEEE Intelligent Systems, pp. 62-68.
[13] D. J. Tritton (1988), Physical Fluid Dynamics, Oxford University Press, New York.
[14] L. Y. Wang, (2000), "Content-based image retrieval for vortex flow data learning environment", master thesis, National Chiao Tung University.
[15] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, (1998) "Adaptive Key Frame Extraction using unsupervised Clustering," Proceedings of Image Processing, ICIP98 , Vol. 1, pp. 866-870.

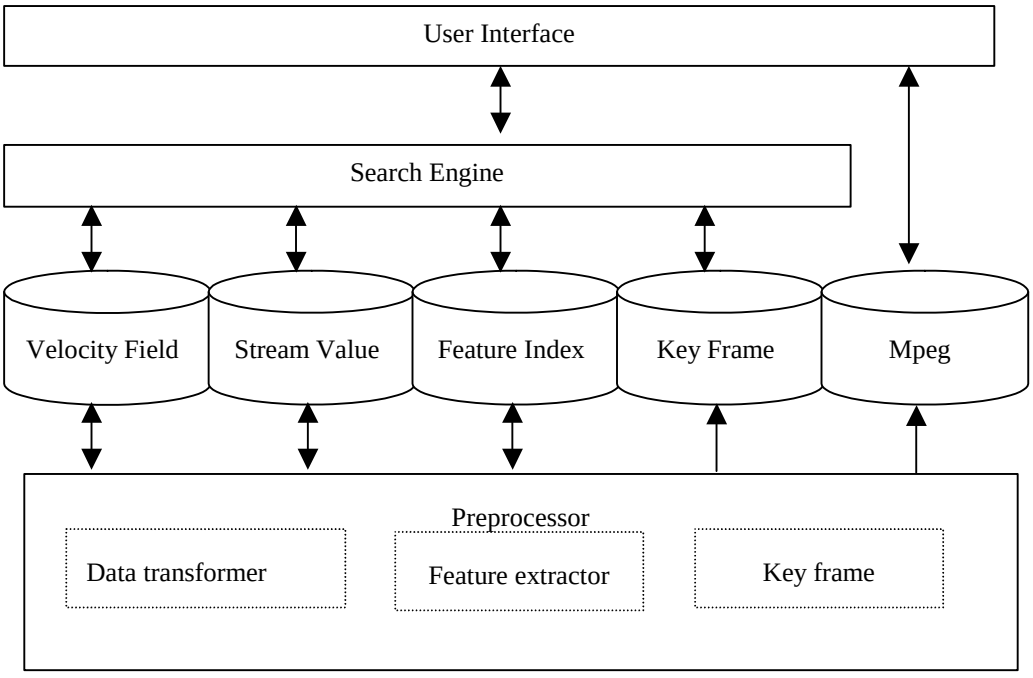

Figure 1: The system flowchart.

Table 1. Preprocessing time.

Velocity Data sets	2000	13000
Stream file generation	19 min.	115 min.
mpeg file generation	57min.	292 min.
Feature extraction	24 min.	158 min.
Key frame extraction	22 sec.	40 sec.

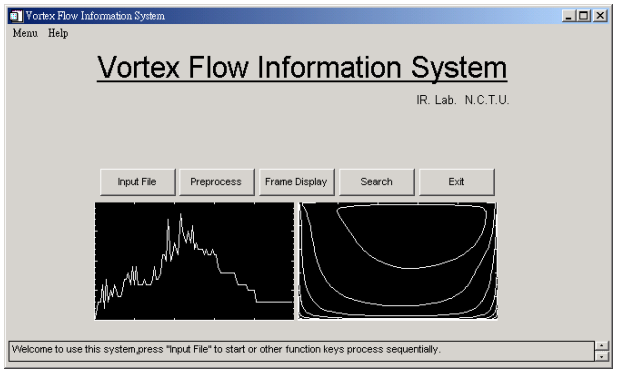

figure 2: welcome page.

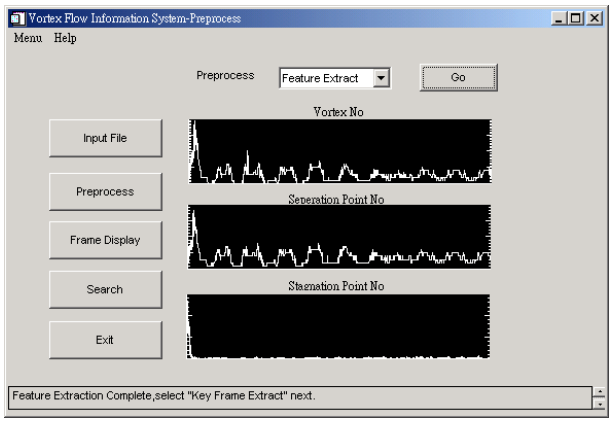

figure 3: feature extraction and results.

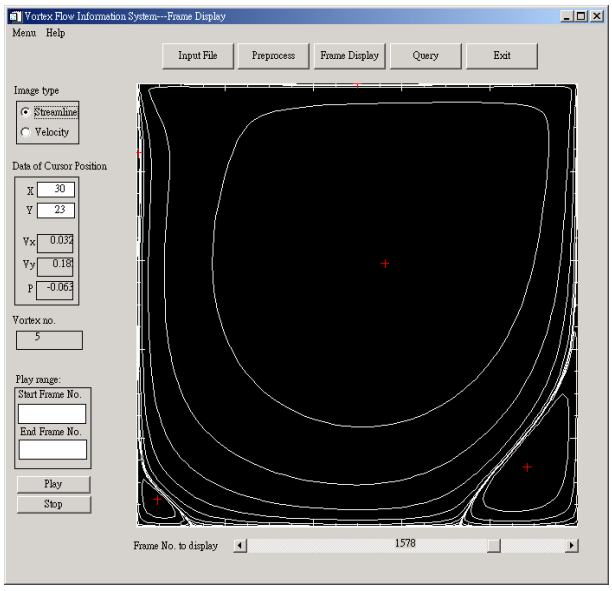

figure 4: frame display and frame record.

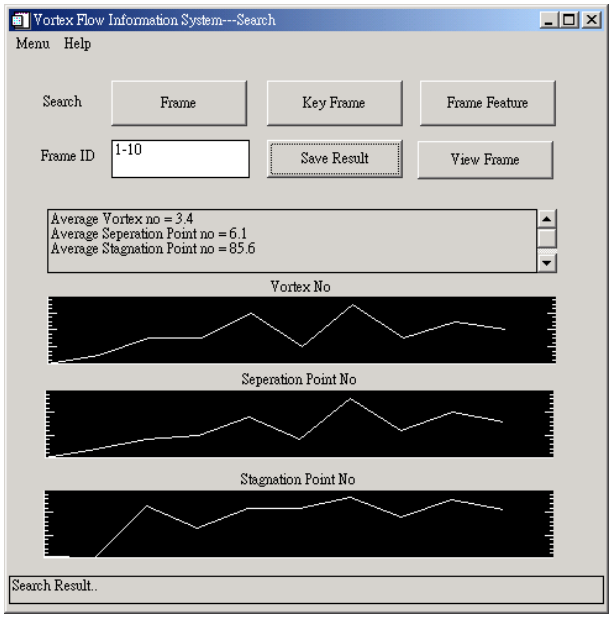

figure 5: frame range search and summary.

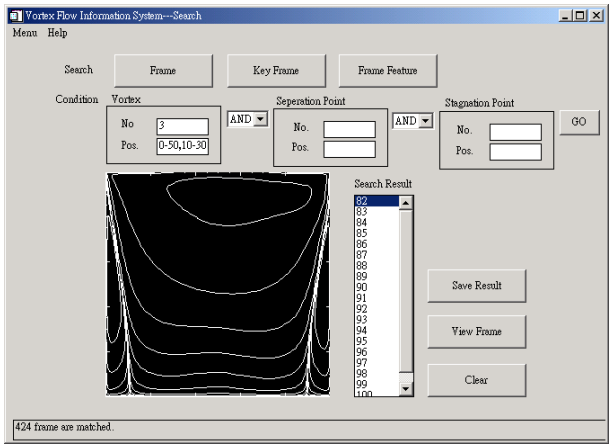

figure 6: Boolean search and its results.

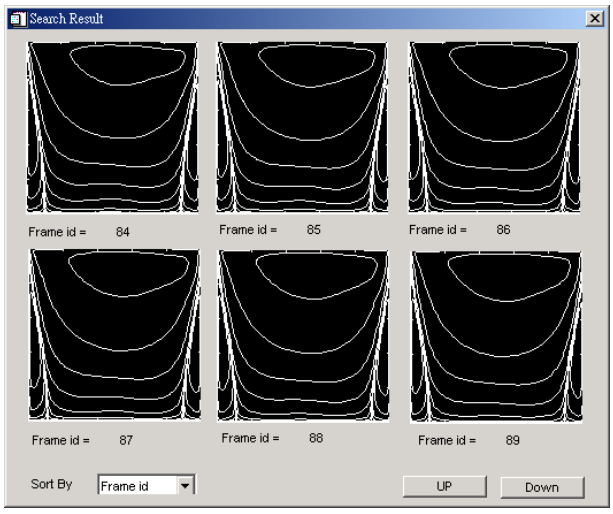

figure 7: query results display.