STATISTICAL APPROACH FOR IMAGE RETRIEVAL

KHOR SIAK WANG

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2007

STATISTICAL APPROACH FOR IMAGE RETRIEVAL

By

KHOR SIAK WANG

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Doctor of Philosophy

January 2007

Family, wife & sons

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

STATISTICAL APPROACH FOR IMAGE RETRIEVAL

By

KHOR SIAK WANG

January 2007

Chairman: Associate Professor Fatimah Bt. Dato' Ahmad, PhD

Faculty: Computer Science And Information Technology

Since the emergence of Internet, a gigantic volume of images have been uploaded into the Internet from time to time. Relying on the traditional text-based search approach to locate the required images could no longer meet the diverse needs of users. This persistent trend has demanded a more sophisticated search algorithm on these images.

One of the popular and common approaches for image search is Content-based Image Retrieval or CBIR for short, i.e. retrieval of images based on their visual contents such as shapes, colours, textures etc.

Of all the visual contents identifiable from an image, colour is considered to be the commonest visual attribute that aids in image retrieval. Works on colour-based image retrieval systems are largely based on the use of colour histogram, which has been noted *to suffer from a major drawback, i.e. absence of spatial information, which is also an important requirement for an accurate retrieval result.*

In this thesis, a novel method based on the modified generic framework of CBIR is proposed. This technique, formally known as Image Retrieval Using Statistical-based Approach is based on the idea of grouping pixels with similar colour codes within an image. From these grouped pixels, they are sorted in descending order of pixel count, which intuitively identifies dominant colours within an image. Statistical information, i.e. means and standard deviations will then be derived from these sorted groups. The extracted statistical information will be stored in both text files and matrixes, which will be used to aid in the image retrieval process. The system has also included some adjustable parameters, such as window size, CC percentage similarity, which can be used to improve retrieval accuracy. This statistical-based approach has been tested on the standard UCID image collection where it has shown improved results, with an average precision value of about 70% as compared to an approximate value of 25% using the histogram-based approach, in term of retrieval accuracy.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

DAPATAN SEMULA IMEJ YANG BERDASARKAN KAEDAH STATISTIK

Oleh

KHOR SIAK WANG

Januari 2007

Pengerusi: **Profesor Madya Fatimah Bt. Dato' Ahmad, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Semenjak kewujudan Internet, terdapat banyak imej yang dimasukkan ke dalam Internet dari semasa ke semasa. Kaedah mendapatkan semula imej secara tradisional yang berdasarkan teks tidak dapat memenuhi keperluan para pengguna. Tren ini memerlukan kaedah pencarian imej yang sopistikated.

Salah satu daripada kaedah yang popular dan biasa untuk mendapatkan semula imej adalah "Content-based Image Retrieval" atau CBIR, iaitu kaedah mendapatkan semula imej berasaskan properti visual seperti bentuk, warna, tekstur dan lain-lain.

Dari semua properti visual yang terkandung di dalam imej, properti warna merupakan properti yang sering digunakan untuk mendapatkan semula imej. Kaedah biasa yang digunakan untuk dapatan semula imej berasaskan warna ialah penggunaan histogram. Kelemahan utama kaedah ini adalah kehadiran lokasi objek di dalam sesuatu imej tidak *dipertimbangkan. Pertimbangan kehadiran lokasi ini merupakan faktor yang penting untuk mendapatkan semula imej dengan tepat.*

Dalam tesis ini, model CBIR yang tradisi akan diubahsuai. Kaedah yang dicadangkan dikenali sebagai Dapatan Semula Imej Yang Berdasarkan Informasi Statistik. Kaedah tersebut berdasarkan idea di mana semua pisel yang mempunyai kod warna yang seragam akan dikelompokkan. Kelompok-kelompok pisel ini akan disusun menurut saiznya. Dengan jelasnya, apabila kelompok tersebut telah disusun mengikut saiznya, ia juga memberi gambaran di mana warna dominan mudah ditentukan. Dari kelompok ini, informasi statistic, iaitu min dan penaburan piawai akan diperolehi. Maklumat tersebut akan disimpan di dalam fail dan array untuk membantu proses dapatan semula imej berasaskan warna. Sistem yang dicadangkan juga mempunyai parameter yang boleh digunakan oleh para pengguna untuk memperbaiki keputusan. Eksperimen yang dilaksanakan dengan menggunakan UCID data dapat menunjukkan kaedah yang dicadangkan mampu memberi keputusan ketepatan secara purata 70% ketepatan dibandingkan dengan 25% dengan menggunakan kaedah histogram.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Associate Professor Fatimah Bt. Dato' Ahmad for giving me an opportunity to start off this project. I 'm indeed obliged by her enthusiastic support of the project from the very early stages. Without her tireless assistance and guidance, this project would never be completed on time. Also, without her constant monitoring and supervisions on the progress of my project, I believe that the contents of this project would still be bits and pieces stored in my hard disk. Her cooperation and contributions are indispensable.

Being a part-time student, I could hardly devote my precious time to my wife, Ms. Kwang Wai Ching, my 3-year old son Khor Hoong Yik and my new-born baby, Khor Hoong Yang, who have been very supportive and patiently waiting for me to complete my study.

Being one of the key persons in the supervisory committee team, Associate Professor Ramlan bin Mahmod is always tight with his schedules and daily events. He really looks serious but approachable. Without his serious-looking face, I would not be able to ensure my work is of the required quality and standard. Thanks, once again.

Associate Professor Hamidah bt. Ibrahim, who is delightful to work with, and always replies me with very short mail on my requests but straight to the point, has been helpful in giving me concrete and constructive comments of my work. I would like to gratefully acknowledge her contributions and her immense help and vast knowledge in database.

Finally, many thanks also go to some of my peer colleagues, where they prefer themselves not to be named, who have given me constructive comments and ideas in certain parts of my research work.

I certify that an Examination Committee has met on 26/01/2007 to conduct the final examination of Khor Siak Wang on his Doctor of Philosophy thesis entitled "Image Retrieval Using Statistical-based Approach" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows: -

ALI MAMAT, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

RAHMITA WIRZA, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia

SHYAMALA DORAISAMY, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia

TENGKU MOHD TENGKU SEMBOK, PhD
Professor
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

HASANAH MOHD. GHAZALI, PhD
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee are as follows: -

Fatimah Dato' Ahmad, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ramlan Mahmod, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Hamidah Ibrahim, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

AINI IDERIS, PhD
Professor/Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

KHOR SIAK WANG

Date:

TABLE OF CONTENTS

Page

DEDICATION *ii*
ABSTRACT *iii*
ABSTRAK *v*
ACKNOWLEDGEMENTS *vii*
APPROVAL *ix*
DECLARATION *xi*
LIST OF TABLES *xv*
LIST OF FIGURES *xvii*
LIST OF ABBREVIATIONS *xx*

CHAPTER

I **INTRODUCTION** *22*
Overview *22*
Need for Image Retrieval *24*
Content-based Image Retrieval *25*
Colour-based Retrieval *27*
Problem Statement *27*
Spatial Information *27*
Dimensionality of Histogram *28*
Objectives of the Research *29*
Main Contributions *29*
Importance of the Research *31*
Thesis Organization *32*
Summary *33*

II **LITERATURE REVIEW** *34*
Overview *34*
Retrieval Levels *35*
1st Level *35*
2nd Level *36*
3rd Level *37*
Traditional Image Retrieval Systems *38*
Content-Independent Image Retrieval *38*
Keyword-based Image Retrieval *38*
Current Research Works *41*
Histogram-based Representation *44*
Integrated Spatial Colour Approach *50*
Back Projection *51*
Colour Coherence Vectors *52*
Colour Correlograms *53*
Spectral Covariance and Fuzzy Regions *54*

Fast Image Retrieval Using Colour-Spatial Data 56
Spatial Colour Histogram 57
Spatial-Chromatic Histogram 58
Partition-based Representation 59
Regional Representation 60
Commercial Applications for Colour Retrieval Systems 61
Query by Image Content (QBIC) 62
VisualSEEk 63
Excalibur Visual RetrievalWare SDK 65
Critical Analysis of Various Approaches 66
Summary 69

III METHODOLOGY 70
Overview 70
System Requirements for Experiment Run 71
Standard Dataset 71
UCID 72
Hardware Requirements 72
Software Requirements 73
Test Data 73
Summary of Query Images 76
Mathematical Proof 81
Summary 82

IV SYSTEM DESIGN OF IMAGE RETRIEVAL USING STATISTICAL-BASED APPROACH 83
Overview 83
Generic Framework for Image Retrieval System 83
Proposed Framework for Image Retrieval System 85
Image Preprocessing 87
Image Retrieval Process 87
Image Acquisition 91
Constraint Imposer 93
Feature Extraction (Image Preprocessing) 96
Indexer 99
Matrix Constructor 100
User Interface 106
User Input 106
Query Parser 109
Feature Extraction (Image Retrieval) 109
Extraction of Essential Information 110
Colour Code 114
CC Extractor 118
SI Retriever 119
DC Identifier 121
Similarity Match Analyzer 122

Window Size *123*
Matched CCs *124*
CC Percentage Similarity *124*
Colour Range Size *125*
Image Matching Process *126*
Stage 1 – Deriving the Number of Matched CCs *126*
Stage 2 – Computing Chromatic Difference *126*
Stage 3 – Incorporating Spatial Information *127*
An Example *129*
Refinery Agent *132*
Calculation of Matching Hits *132*
Promote/Demote Process *133*
Ranker *133*
Retrieval Manager *134*
Relevance Feedback *134*
Functional Flow *136*
Descriptions of Flowcharts *141*
Selection Sort *148*
Summary *149*

V RESULTS AND DISCUSSIONS *150*
Overview *150*
Summary of Query 1 (Histogram Approach) *151*
General Information *151*
Adjustable Parameters *151*
Retrieved Images *152*
Possible Similar Images (NOT in ranked order) *153*
Summary of Query 1 (Proposed Technique) *155*
General Information *155*
Adjustable Parameters *155*
Retrieved Images *156*
Possible Similar Images (NOT in ranked order) *157*
Discussions of Results for Query 1 *160*
Conclusion *160*
Summary of Results *161*
Overall Conclusions and Findings *169*
Scenario-based Illustrations *169*
Summary of Individual Results *170*
Overall Summary of Results *175*
Overall Conclusion and Findings *177*
Reduced Computational Load *177*
Summary *180*

VI CONCLUSION AND FUTURE WORKS *181*
Conclusion *181*
Future Works *182*

Summary *184*

REFERENCES/BIBLIOGRAPHY *185*
APPENDICES *192*
BIODATA OF THE AUTHOR *197*

LIST OF TABLES

Table **Page**

3.0 *Sample Query Images* *79*

3.1 *A Summary of Adjustable Parameters* *80*

3.2 *Five Unique Scenarios* *82*

4.0 *A Summary of Roles and Functions for Each Subsystem in the Proposed Model for Colour Image Retrieval System* *90*

4.1 *Structure of the Text File, IndTbl.txt* *99*

4.2 *Default Values of the Adjustable Parameters* *109*

4.3 *CC Values* *118*

4.4 *A Summary of Images for Image Matching Process* *129*

4.5 *A Summary of CCs and Pixel Counts for Both Query Image and Stored Image* *130*

4.6 *Arrays Used in Image Retrieval Process* *144*

5.0 *Histogram – Details of Query Image (Query 1)* *151*

5.1 *Recall & Precision Values for Query 1 (Histogram)* *154*

5.2 *Details of Query Image (Query 1)* *155*

5.3 *Recall & Precision Values for Query 1* *158*

5.4 *Summary of Findings (Query 1)* *159*

5.5 *Interpolated Results (Query 1)* *159*

5.6 *Summary Findings of Histogram Approach* *164*

5.7 *Summary Findings of the Proposed Technique* *165*

5.8 *Averaged Precision Values for Thirty Queries* *167*

5.9 *Statistical Information for Scenario 1 (Query Image)* *170*

5.10 Statistical Information for Scenario 1 (Stored Image) 170

5.11 Statistical Information for Scenario 2 (Query Image) 171

5.12 Statistical Information for Scenario 2 (Stored Image) 171
5.13 Statistical Information for Scenario 3 (Query Image) 172

5.14 Statistical Information for Scenario 3 (Stored Image) 172

5.15 Statistical Information for Scenario 4 (Query Image) 173

5.16 Statistical Information for Scenario 4 (Stored Image) 173

5.17 Statistical Information for Scenario 5 (Query Image) 174

5.18 Statistical Information for Scenario 5 (Stored Image) 174

5.19 Summary of Statistical Information 176

LIST OF FIGURES

Figure **Page**

1.0 A General Problem of Image Retrieval 22

2.0 Some Sample Images for Retrieval at 1st Level 36

2.1 Some Sample Images for Retrieval at 2nd Level 36

2.2 Some Sample Images for Retrieval at 3rd Level 37

2.3 A Sample Query on Keyword-based Image Search
Engine 39

2.4 General Approaches for Colour-based Image Retrieval
Systems 43

2.5 A Flying Eagle 44

2.6 Histogram 44

2.7 Two Images With Simple Objects 47

2.8 Identified Problems in Histogram-based Colour Retrieval
System 49

2.9 Sample Image Where Object is Placed at Center 55

2.10 Distribution State of Pixels 57

2.11 An Image Being Divided Into six Partitions 59

2.12 An Image Composed of Salient Regions 60

2.13 A Red Rose 61

2.14 Sample Query Screen of QBIC 63

2.15 Image Decomposition 63

2.16 A Sample Input Screen of VisualSEEk 65

2.17 A Sample Input Screen of Excalibur Visual RetrievalWare 66

4.0 A Generic Framework for the Color Image Retrieval
System 84

4.1 A Proposed Framework for the Color Image Retrieval System 86

4.2 Constraints on Pre-processed Image 93

4.3 Structure of the Text File, IndTbl.txt 100

4.4 Arrays That Hold Data of the Text File 101

4.5 Schematic View of Image Preprocessing 105

4.6 Main Menu 106

4.7 Image Preprocessing 107

4.8 Query Screen 107

4.9 Search Parameters 108

4.10 Sample 1 - Stored And Query Images 111

4.11 Sample 2 - Stored And Query Images 111

4.12 Sample 3 - Stored And Query Images 112

4.13 Discretization of the RGB Components 116

4.14 Sample Image With its Distribution State of Pixels 120

4.15 Window Size 123

4.16 Extracted Information of the Query Image 129

4.17 Extracted Information of the Stored Image 130

4.18 Flowchart Showing Image Preprocessing (for Stored Images) 137

4.19 Flowchart Showing Functional Flow of the Retrieval Process 138

4.20 Flowchart Showing Functional Flow of the Feature Extraction Process (for Query Image) 139

4.21 Flowchart Showing the Image Matching Process 140

4.22 Manner in Which Pixels are Being Processed 141

5.0 Retrieved Images (First Page) – Query 1 (Histogram) 152

5.1 Retrieved Images (Second Page) – Query 1 (Histogram) 152

5.2 Retrieved Images – Query 1 (Histogram) 153

5.3 Retrieved Images (First Page) – Query 1 156

5.4 Retrieved Images (Second Page) – Query 1 156

5.5 Retrieved Images – Query 1 157

5.6 Recall & Precision Graph (Query 1) 160

5.7 Results of the Query Image "Flag" 162

5.8 Results of the Query Image "Dancers" 163

5.9 Averaged Recall & Precision Values (Line Chart) 168

5.10 Averaged Recall & Precision Values (Bar Chart) 168

LIST OF ABBREVIATIONS

2D	*Two Dimensions*
3D	*Three Dimensions*
ATM	*Asynchronous Transmission Mode*
CAD	*Computer-aided Design*
CBIR	*Content-based Image Retrieval*
CBVIR	*Content-based Visual Information Retrieval*
CC	*Colour Code*
CCV	*Colour Coherence Vector*
CD	*Compact Disk*
CIE	*Commission Internationale de l'Êclairage*
CMY	*Cyan (C), Magenta (M), and Yellow (Y)*
CRT	*Cathode Ray Tube*
DC	*Dominant Colour*
FE	*Feature Extraction*
GUI	*Graphical User Interface*
HIS	*Hue-Intensity-Saturation*
HSV	*Hue-Saturation-Value*
IR	*Information Retrieval*
ISDN	*Integrated Services Digital Network*
MIR	*Multimedia Information Retrieval*
MARS	*Multimedia Analysis and Retrieval System*
MPEG	*Moving Picture Experts Group*

QBIC	*Query By Image Content*
RGB	*Red-Green-Blue*
RF	*Relevance Feedback*
SCH	*Spatial-Chromatic Histogram (SCH)*
SI	*Statistical Information*
SMAT	*Sequenced Multi-Attribute Tree*
SONET	*Synchronous Optical Network*
SQL	*Structured Query Language*
UCID	*Uncompressed Colour Image Database*
VLSI	*Very Large-Scale Integration*