多媒體資訊檢索

王啟正
中央研究院 資訊科技創新研究中心

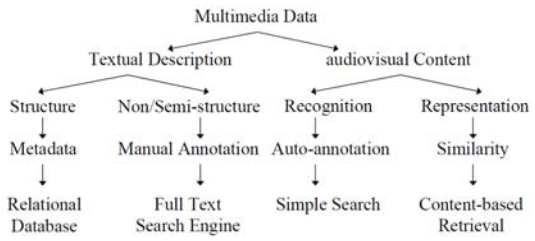
VCenter (http://vcenter.iis.sinica.edu.tw) iPicBox.tw

前言

- **多媒體**泛指多種媒體型態的融合，包含文字(text) 、圖像(image) 、影像(video) 、語音(speech) 、音樂(music) 、動畫(animation)等等
- **資訊檢索**是擷取、組織和利用資訊的重要技術

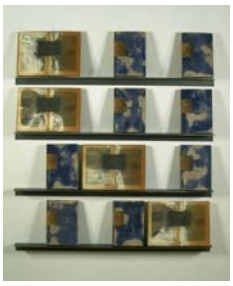
Data is of no use
unless you can actually access it.

多媒體資訊檢索系統 概略圖



Similar?

Similar?



Similar?

Similar?

7

Similar?

8

Similar?

9

Similar?

10

Show Demo

- **ACM Multimedia 2010 - Technical Demonstrations, Paper ID: 1569322929**

Outline

- 多媒體資料庫搜尋
- 多媒體全文搜尋引擎
- 以內容為基礎之多媒體檢索
- 多媒體資訊系統應用

http://ebook.iis.sinica.edu.tw/

多媒體資料庫搜尋 以CalPhotos為例

http://calphotos.berkeley.edu/

Multimedia Metadata

- Bibliographic attributes
- Sensory perception
- Generic knowledge

Bibliographic attributes

- DC (Dublin Core)
- CDWA (Categories for the Description of Works of Arts)
- EAD (Encoded Archival Description)
- CSDGM (Content Standards for Digital Geospatial Metadata)
- EXIF (Exchangeable Image File Format)

Example: EXIF

- 可交換影像檔案格式(Exchangeable Image File Format)
- 日本電子工業發展協會(Japan Electronic Industry Development Association, JEIDA)
- 1996年制定，版本為1.0
- 1998年升至2.1
  - 增加對影音檔案的支援
- 2002年發表2.2版
  - EXIF可以附加於JPEG、TIFF、RIFF等檔案類型的標頭檔(Header)裡
  - 用來描述數位相機圖檔的相關資訊，如：相機品牌、型號、拍攝日期、光圈、ISO值、白平衡…

可讀取EXIF軟體:
- Opanda IExif 2.25
  - http://www.opanda.com/
- Exif Reader

屬性名稱（英文）	屬性名稱（中文）	屬性值
主要資訊		
Make	製造商	FUJIFILM
Model	型號	FinePix4700 ZOOM
Orientation	影像方向	Left-hand side
X Resolution	水平解析度	72/1
Y Resolution	垂直解析度	72/1
Resolution Unit	解析度單位	Inch
Software	韌體版本	Digital Camera FinePix4700 ZOOM Ver. 1.00
Date Time	拍攝時間	2000:01:01 13:39:44
YCbCr Positioning	YCbCr位置	co-sited
Copy Right	版權	
Exif Info Offset	Exif資訊偏移量	262
次要資訊		
F Number	光圈(F)值	F2.8
Exposure Program	曝光程序	Program
ISO Speed Ratings	ISO值（感光度）	200
Exif Version	Exif版本	0210
Date Time Original	原始成像時間	2000:01:01 13:39:44
Date Time Digitalized	相機存檔時間	2000:01:01 13:39:44
Component Configuration	Component Configuration	YCbCr
Compressed Bits Per Pixel	壓縮後每像素位元數	2/1 (bit/pixel)
Shutter Speed Value	快門速度	1/49Sec
Aperture Value	光圈值	F2.8
Brightness Value	亮度	EV2.8
Exposure Bias Value	曝光偏移	EV0.0

屬性名稱（英文）	屬性名稱（中文）	屬性值
Max Aperture Value	最大光圈值	F2.8
Metering Mode	測光模式	Division
Flash	閃光燈	Not fired
Focal Length	焦距	8.30(mm)
Marker Note	Marker Note	FUJIFILM Format : 214Bytes (Offset:732)
Flash Pix Version	FlashPix版本	0100
Color Space	色彩空間	sRGB
Exif Image Width	圖像寬度	1280
Exif Image Height	圖像高度	960
Exif Interoperability Offset	Exif Interoperability Offset	938
Focal Plane X Resolution	水平原始解析度	1270/1
Focal Plane Y Resolution	垂直原始解析度	1270/1
Focal Plane Resolution Unit	原始像素單位	Centimeter
Sensing Method	感光元件	One Chip Color Area sensor
File Source	檔案來源	DSC
Scene Type	取景方式	A directly photographed image
廠商原始資訊		
Version	版本	0130
Quality Mode	圖像品質	
Sharpness	銳利度	
White Balance	白平衡	Auto
Flash Mode	閃光燈模式	Off
Flash Strength	閃光時間	0/10

屬性名稱（英文）	屬性名稱（中文）	屬性值
Macro Mode	近拍模式(?)	On
Focus Mode	聚焦模式	Auto Focus
Focus Mode	聚焦模式	Auto Focus
Slow Sync	慢速同步	Off
Mode	模式	Auto
Sequence Mode	序列模式	Off
Blurring Warning	模糊警示	Yes
Focus Status	聚焦狀態	Nice
Exposure Status	曝光狀態	Nice

圖像Metadata

- 內容資訊
  - 除了圖像的EXIF資訊以外，一般人都會想知道關於這張圖的內容。因此可以透過Metadata的制定，針對圖像進行詳細的描述。
  - 如右圖
    - 相片標題
    - 相片資訊
    - 相片標籤

Sensory perception

- Sensory perception is about the low-level feature representation of multimedia data(color, texture, shape, audio, motion, ...)
- We will discuss it in content-based retrieval

Generic knowledge

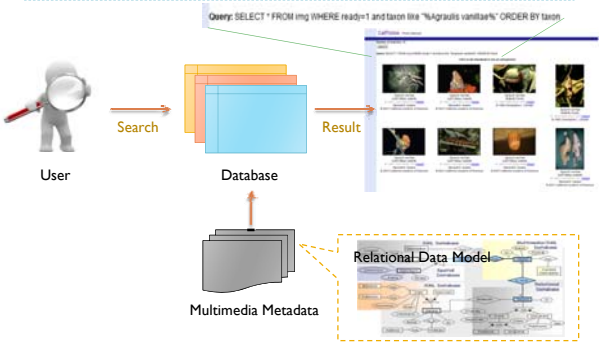
- High-level semantic features (subject, of/about, generic/specific, time, space, activity, object,…)

層面（Facets）	屬於（Of）		關於（About）	非層面專指的相關意義
	狹義（Specific Of）	廣義（Generic Of）		
人物（Who）	有個別名稱的人、動物、事物	某一類型的人、動物、事物	某人或物所象徵的抽象意涵	
事件（What）	有個別名稱的事件	行動、情況	情緒或事件的抽象表達	
地點（Where）	有個別名稱的地理區	某一個地理區或建築地	含有抽象象徵意義的地區	
時間（When）	某個時間、日期或時期	某個季節或時日	某個時間所象徵的抽象情緒	

- Perception subjectivity & labor-intensive annotation

Sara Shatford,"Analyzing the subject of a picture: a theoretical approach," Cataloging & Classification Quarterly 6 : 3 (1986): 49-54.

多媒體資料庫搜尋



Outline

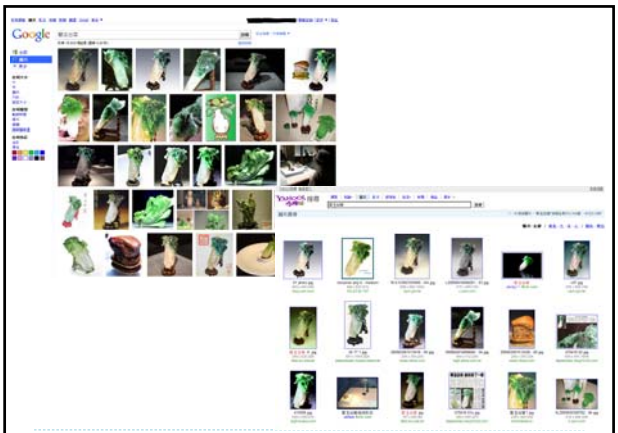
- 多媒體資料庫搜尋
- 多媒體全文搜尋引擎
- 以內容為基礎之多媒體檢索
- 多媒體資訊系統應用

http://ebook.iis.sinica.edu.tw/

多媒體全文搜尋引擎

- **全文搜尋**為從文本或資料庫中，不限定資料欄位，自由地萃取出訊息的技術。
- 目前常見的多媒體全文搜尋引擎
  - Google
    - Google 圖片 | Google 影片
  - YAHOO!
    - YAHOO!圖片 |YAHOO!影片
  - 其它
    - blinkx

您只要在Google search上輸入"image/video/music seach engine"關鍵字，就可以找到各式各樣的檢索服務網站

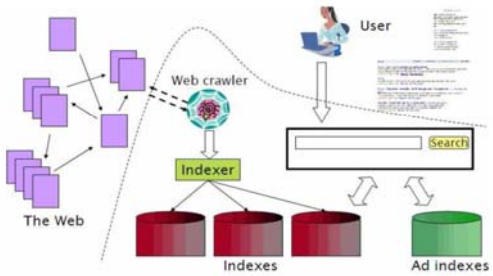


http://catalog.digitalarchives.tw/

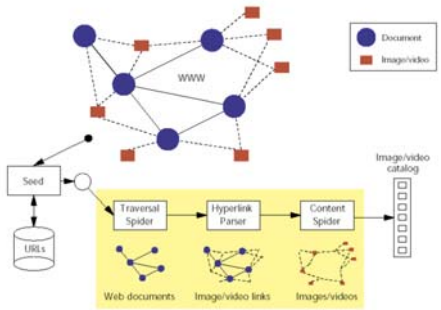
Textual Multimedia

- Images of documents that contain mainly typed or typeset text
- How to obtain?
  - Use a set of keywords, i.e., **metadata**, associated with each multimedia at creation time
  - Use OCR to recognize the words in the multimedia
  - Use the **web** crawler to extract texts around the multimedia

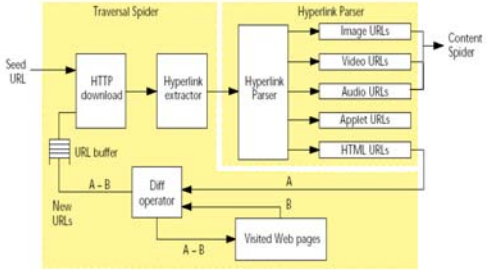
Web Search Overview



Gathering Web Multimedia File



Gathering Web Multimedia File



Semantic Extraction from Web

- From Web Pages
  - Filename/URL, anchor text, image caption, page title, alternative text (ALT = ...), surrounding text
- From multimedia file headers
  - File type, file size, EXIF attributes, MPEG attributes, MP3 attributes

Recall vs. Precision

- Recall（回收率/查全率）
  - Precision = Relevant Documents Retrieved / Total Retrieved Documents
- Precision（精確率/查準率）
  - Recall = Relevant Documents Retrieved / Total Relevant Docuements

	相關	不相關
檢索到	a	b
未檢索到	c	d
總數	a+c	b+d

$$\text{回收率} = \frac{a}{a+c} = \frac{\text{檢索所得之相關文章筆數}}{\text{資料庫中所有相關文章筆數}}$$

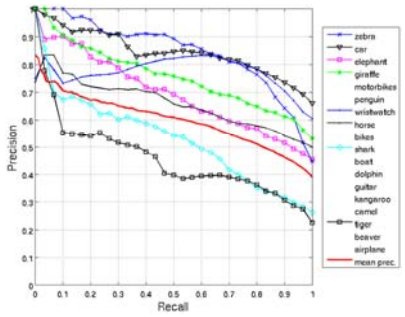
$$\text{精確率} = \frac{a}{a+b} = \frac{\text{檢索所得之相關文章筆數}}{\text{檢索所得之所有書目筆數}}$$

Example:

- 舉個例子：假設現在資料庫中有10000筆資料，和旅遊有關的文章有500篇。使用者在輸入旅遊的關鍵字後，回傳的文章有4000篇，其中有400篇是和旅遊有關的。

查準率 Precision = 400 / 4000 = 10%
查全率 Recall = 400 / 500 = 80%

Recall vs. Precision (Example)



Harvesting Image Databases from the Web
http://www.robots.ox.ac.uk/~vgg/research/mkdb/

多媒體全文檢索

- Text-Based Multimedia Retrieval
  - Uses text descriptions to retrieve relevant multimedia
  - Time, location, events, objects
  - Advantage – based on currently in use text retrieval system
  - Disadvantage
    - Manual annotation is **labor sensitive** and **subjective interpretation**
    - An image is worth a thousand words?

But the fact is a few words labeled!

Outline

- 多媒體資料庫搜尋
- 多媒體全文搜尋引擎
- 以內容為基礎之多媒體檢索
- 多媒體資訊系統應用

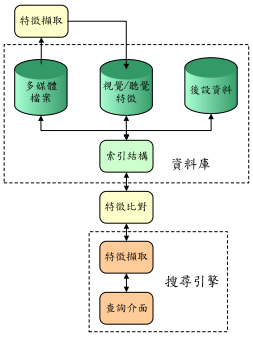
http://ebook.iis.sinica.edu.tw/

以內容為基礎之多媒體檢索

- Content-Based Multimedia Retrieval
  - Multimedia files that are indexing and retrieval based on audiovisual perceptual features, which can be extracted automatically
    - **Image:** color, texture, shape, spatial layout
    - **Video:** motion, image, audio
- Extracts images based on image content.
  - Level 1: Retrieval by primitive features such as color, texture, shape and spatial location.
  - Level 2: Retrieval of objects of given type. Example: find the picture of the flower.
  - Level 3: Retrieval of abstract attributes that involves high level reasoning. Example: 'find picture of a baby smiling'.

General Framework

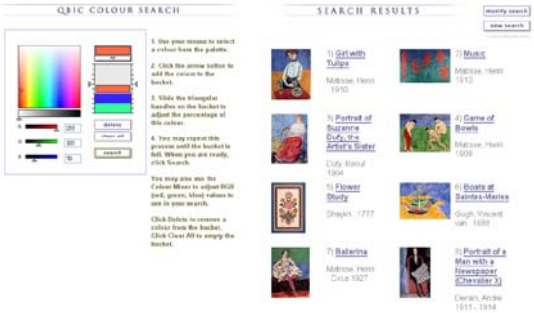
- **Query interface**
  - Query type support
- **Browsing**
  - Data collection Navigation
- **Feature extraction**
  - Data collection Navigation
- **Indexing**
  - Similar to the IR model
- **Similarity measurement**
  - Similar to the IR model
- **Relevance feedback**
  - Similar to the IR model



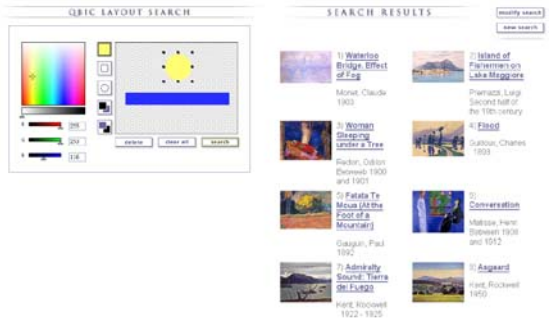
Query Interface in Content-based Retrieval

- Direct query (or concept query)
  - Use user interface to create a query (i.e., sketches, drawing, etc.)
- Query by example
  - Provide a multimedia document as a query
- Relevance feedback
  - Select multiple relevant or irrelevant examples from the search result

Direct Query – QBIC Color Search (IBM)

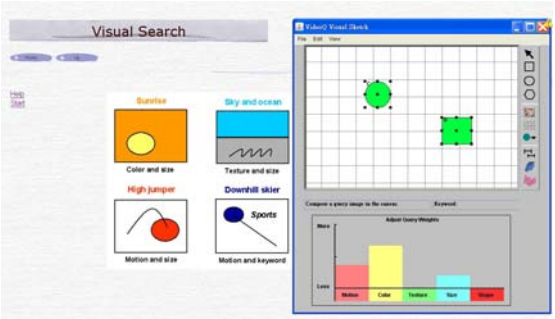


Direct Query – QBIC Layout Search (IBM)

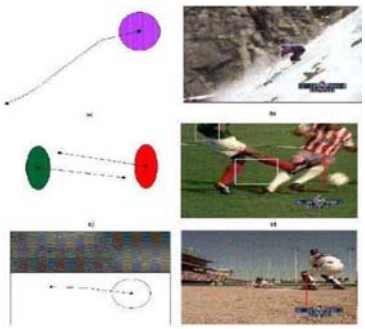


Direct Query – CSI iPicBox.tw

Direct Query – VideoQ (Columbia University)



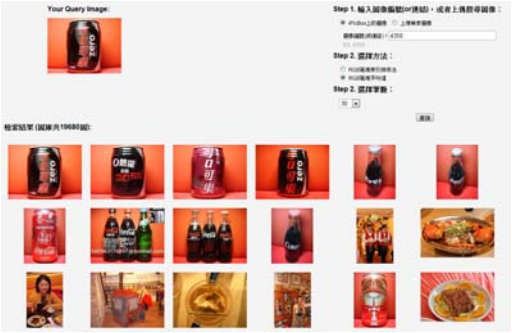
Direct Query – VideoQ (Columbia University)



Query by Example – SIMPLIcity (Stanford University)

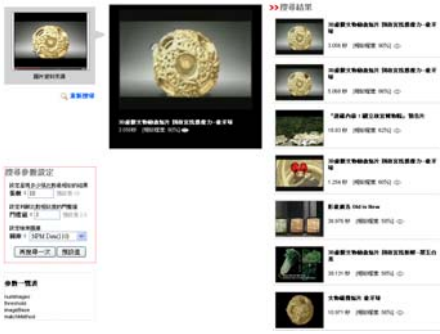
http://wang14.ist.psu.edu/cgi-bin/zwang/regionsearch_show.cgi

Query by Example – ISIC iPicBox.tw



Query by Example – dwtsi iPicBox.tw

Query by Example – VCenter



http://vcenter.iis.sinica.edu.tw/lab/shotDetect/camera.html

Query by Spoken Content

- 中研院資訊所王新民副研究員發展的語音辨識技術，可以用來檢索影片中的語音內容

Relevance Feedback – 新視科技

Relevance Feedback – 新視科技

Relevance Feedback – 新視科技

Browse Multimedia Content

- Subject browsing is the most popular user operation in multimedia retrieval interactive
- Browsing is helpful when ...
  - Navigating the multimedia collection
  - Specifying an example as the query
  - Comparing the retrieval result

A Case Study: WebSEEk

- An image and video search engine for the Web
  - Information is collected by automated agents
  - Text and visual features are both processed
  - Media objects are classified
- Using URLs, alt tags and hyperlink text to create subject taxonomy (2128 classes)
- Using color histograms to classify image types
  - Color photo, color graphic, gray image, B/W image

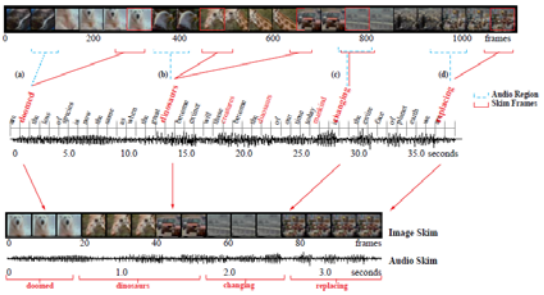
Image Taxonomy in WebSEEk

Video Browsing

The Open Video Project

56

影片摘要（video summarization）



M. Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization", *Technical Report CMU-CS-95-186, School of Computer Science, Carnegie* Mellon University, July 1995.

Feature Extraction

- **Image**
  - Color
  - Texture
  - Shape
  - Spatial layout
- **Video**
  - Motion
  - Image (Frame, Shot)
  - Audio

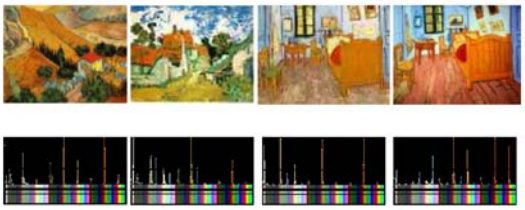
MPEG-7

- **Multimedia content description interface**
- **Standardize tools** (descriptors, description schemes, and language) of audiovisual information
- Enable fast and efficient **content searching, filtering, and identification**
- Proposed by ISO/IEC

Color

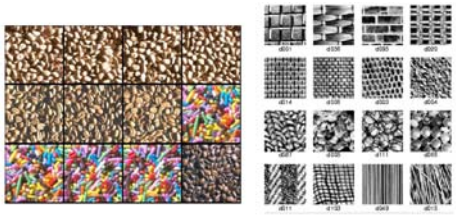
- Color Characteristics of images
- Selection of Color Space is important in CBIR
  - Examples → RGB, CMY
- Averaging colors for regions
- Histograms
- Vectors

Color Histogram



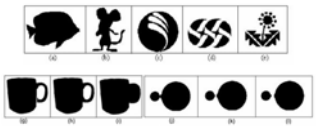
Texture

- Visual Pattern
- Uniform intensity region of a simple shape that is repeated
- The structural and statistical properties of brightness pattern



Shape

- Involves detecting the border and boundary of objects.
- This is done by edge detection algorithms.
- They measure area, circularity, shape signature, curvature, eccentricity to determine the shape.
- Different objects such as flowers, brain tumors etc require different algorithms. Hence shape preprocessing algorithms are application dependent.



Spatial Location

- This feature is used for region classification.
- Spatial location are defined as 'upper, bottom, top' according to the location of the region in the image.
  - For example sea and sky could have similar color and texture features, but their spatial locations are different.

Points of Interest

- SIFT (Scale-Invariant Feature Transform)
  - Detect scale-space extrema points (scale invariant)
  - Localize keypoints by filtering out low contrast points
  - Assign major orientations for each keypoint (rotation invariant)
  - Calculate the orientation histogram for each keypoint to form a 128-dimensional vector (illumination and viewpoint invariant)
- CCH (Contrast Context Histogram)
  - Similar to SIFT
  - IIS, Sinica

65

影音內容分析

- Motion Feature
- 鏡頭變化偵測(Shot detection)
- 字幕辨識(Caption recognition)
- 語音辨識(Speech recognition)
- 人臉偵測與人臉辨識(Face detection & Face recognition)

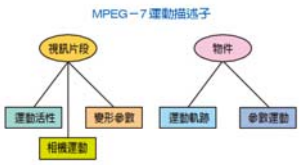
66

Motion

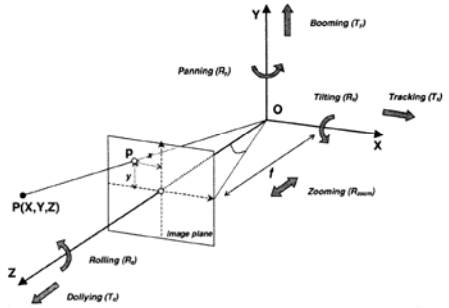
- The motion features of a video sequence provide the easiest access to its **temporal dimension**
- In combination with indexing based on still-image
- features, motion-based indexing is useful in contexts where motion has a **higher-level semantic meaning**

Motion Feature

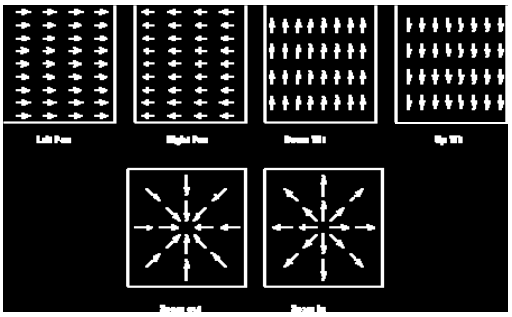
- Global motion
  - Motion activity
  - Camera motion
  - Warping parameters
- Local motion
  - Motion trajectory
  - Parametric motion



Camera Operation

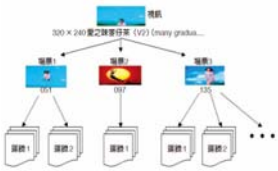


Optical Flow of Camera Motion



Shot

- 視訊內容最基本的單位是Frame ，它相當於一張影像，但在語意(semantics)上並不具太大的意義
- 在視訊內容中具有語意的最基本單位是shot
  - Shot, 為影片在拍攝過程中, 導眼命令攝影師打開攝影機開拍一直到關機停拍這一段期間



鏡頭變化

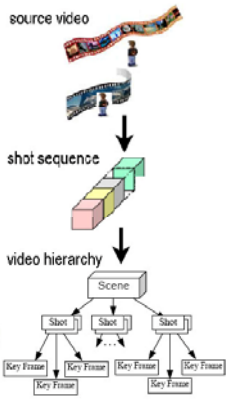
- 突然式的鏡頭變化(跳接鏡頭) Cut (abrupt transition)

- 淡入/淡出鏡頭 Fade in/out (gradual transition)

- 溶解鏡頭 Dissolve (gradual transition)

鏡頭偵測

- 偵測影片中畫面變換的時間點 (shot detection)
- 擷取每一分鏡片段中的一小段影片，組合成該影片之動態摘要片段。
- 擷取每一分鏡片段的關鍵畫格(key frames)，作為該影片之靜態摘要圖像。

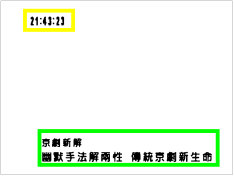


73

C. C. Shih, H. Y. M. Liao, and H. R. Tyan, "Shot change detection based on the Reynolds transport theorem," *IEEE Pacific Rim Conference on Multimedia*, 2001.

字幕辨識

- 辨識影格(frame)中的文字(如字幕)可供文字檢索(text-based retrieval)的索引資訊
- 目前字幕辨識的辨識率:中文字幕可達90% (640x480), 英文字幕可達70%(320x240)

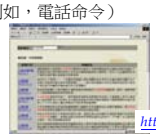


74

F. Chang, G. C. Chen, C. C. Lin, and W. H. Lin, "Caption analysis and recognition for building video indexing system," *ACM Multimedia Systems*, 2005.

語音辨識

- 語音辨識目的是將語音訊號自動轉換成文字描述，提供內容文字檢索 (text-based retrieval)
- 針對特定演講者所進行的語音辨識，且滿足以下的條件，其辨識率可達 98%
  - 事先已針對該名演講者的語音資料進行訓練， 取得其語音 訊號的特徵值資訊
  - 適當地調整演講者的聲調高低
  - 安靜的環境底下
- 對於有限的字彙，系統可以不用透過語意訓練就能夠辨識到大多人的所說的少數字詞（例如，電話命令）

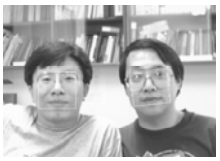


http://sovideo.iis.sinica.edu.tw/index.html

H. M. Wang, S. S. Cheng, and Y. C. Chen, "The SoVideo Mandarin Chinese Broadcast News Retrieval System," *International Journal of Speech Technology*, 2004.

人臉偵測與人臉辨識

- 人臉偵測與人臉辨識技術可從影片中搜尋特定人物作為索引資訊
  - 例如尋找美國布希總統與英國首相布萊爾握手的畫面



根據科學相關的論文和報告，人臉偵測精確度可達到90 %

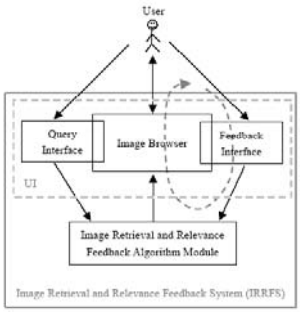
人臉偵測與人臉辨識

Picasa

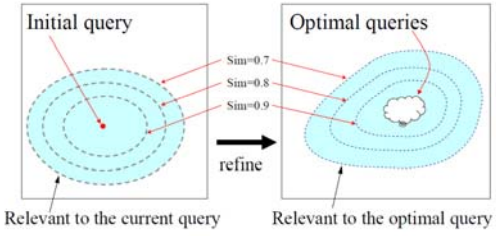
- Picasa 3.6 的功能

新增名稱標記：Picasa 會掃描圖片集中的所有相片，並找出包含臉孔的相片，再根據臉孔相似度將相片分組。如要同時為數十張相片新增名稱標記，只要按一下相片下方的 [新增名稱]，然後輸入該位人物的名稱即可。新增名稱標記之後，您可以透過簡單的單鍵操作方式建立臉孔美術拼貼、輕鬆搜尋包含兩位相同人物的相片，或是將您的名稱標記上載到「Picasa 網路相簿」。

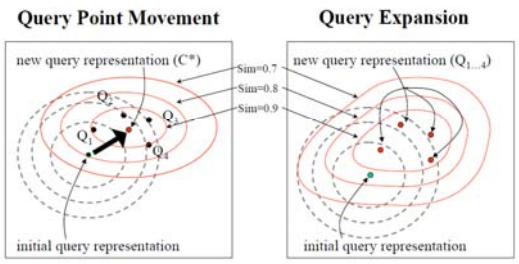
Relevance Feedback Interface



Relevance Feedback



Relevance Feedback



Outline

- 多媒體資料庫搜尋
- 多媒體全文搜尋引擎
- 以內容為基礎之多媒體檢索
- **多媒體資訊系統應用**

http://ebook.iis.sinica.edu.tw/

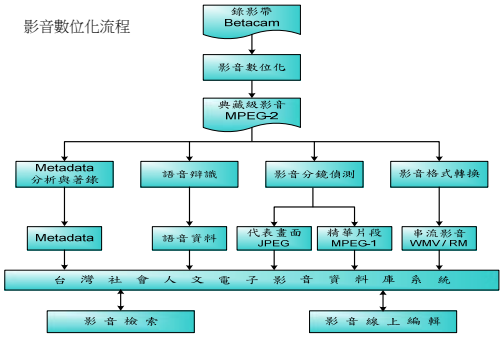
台灣社會人文影音資料庫

- 以研究電子影音資料典藏為目標的應用性研究推廣計畫
- 典藏的內容
  - 國立臺北藝術大學科技藝術研究中心的李道明教授十多年來，從事臺灣社會與人文相關紀錄片為基礎，免費提供民眾查詢使用
- 以ECHO metadata規範為基礎進行修改，以符合計畫之需求

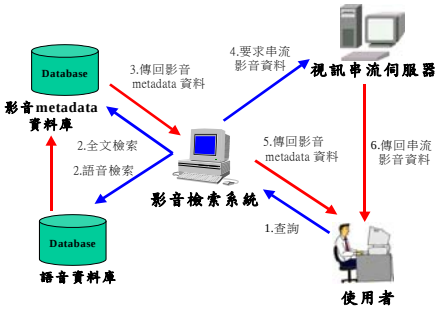
http://ec-video.iis.sinica.edu.tw/

台灣社會人文影音資料庫

影音數位化流程



影音檢索流程範例



影音資料檢索系統畫面

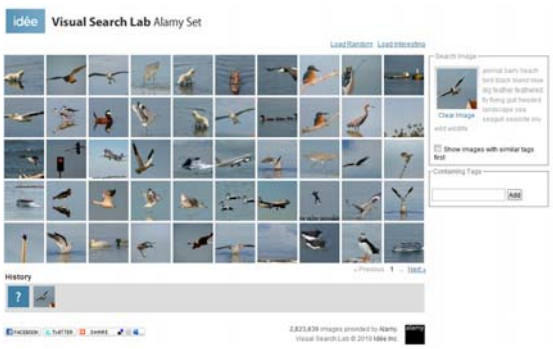
TinEye

- Idée, Inc.公司所開發的圖像搜尋軟體。
- 2008年5月6日開始提供使用者依據連結或上傳的圖檔搜尋網路上類似的圖像位置。
- PixID軟體，分辨影像或靜態圖像是否在網際網路中出現。
- Piximilar將類似影像或靜態圖像檔分類，並建立索引檔管理圖像檔案。
- 簡介影片

http://ideeinc.com/products/pixid/

Idée Labs

http://labs.ideeinc.com/multicolr/



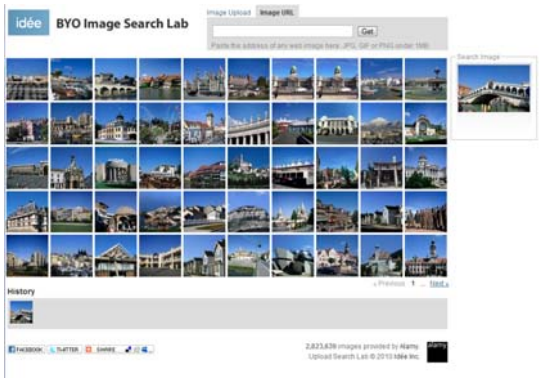
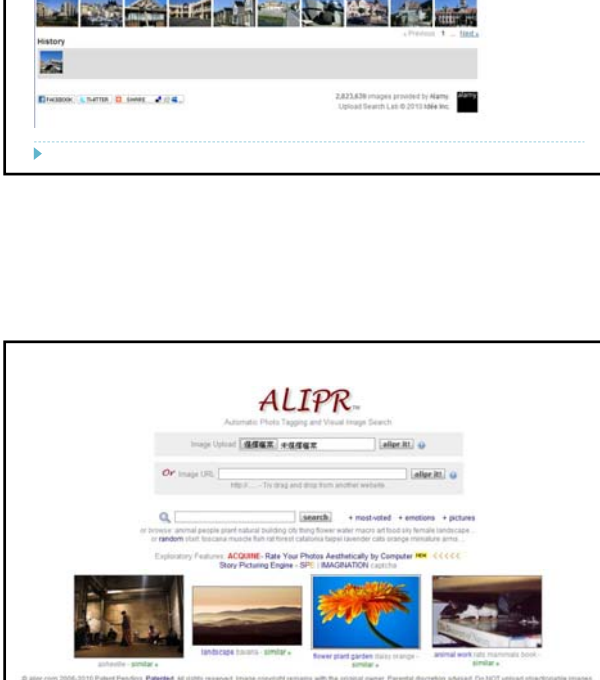
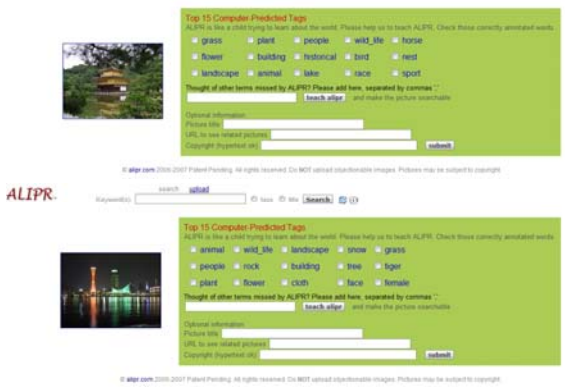


Image Tagging

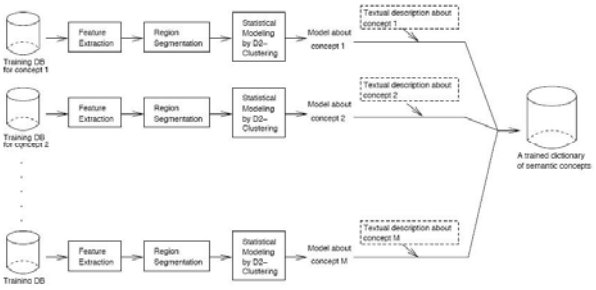
- Manual
  - Advantage- more accurate, custom tags
  - Disadvantage- tedious, human error, not always tagged
- Automatic
  - **Image annotation**
  - Advantage- quick & easy, every image is tagged
  - Disadvantage- limited, inaccurate



ALIPR – Automatic Linguistic Indexing of Pictures in Real Time http://alipr.com/



ALIPR Diagram



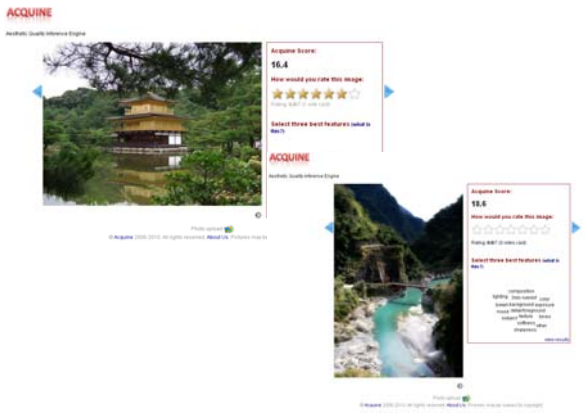


Image-based Rendering (IBR)

- IBR relies on a set of two-dimensional images of a scene to generate a three-dimensional model and then render some novel views of this scene.
- Web 2.0 enables sharing of photographs on a truly massive scale

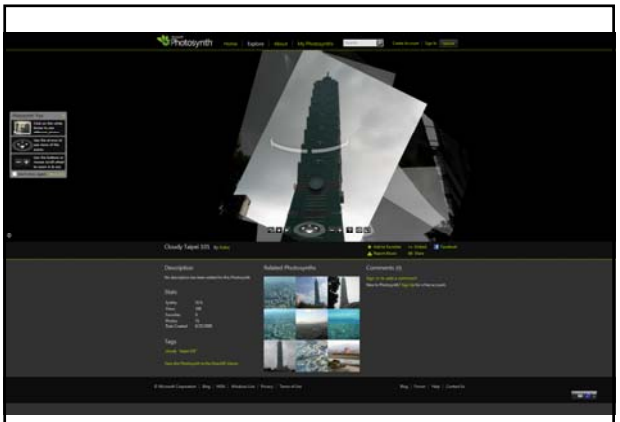
98

Microsoft PhotoSynth

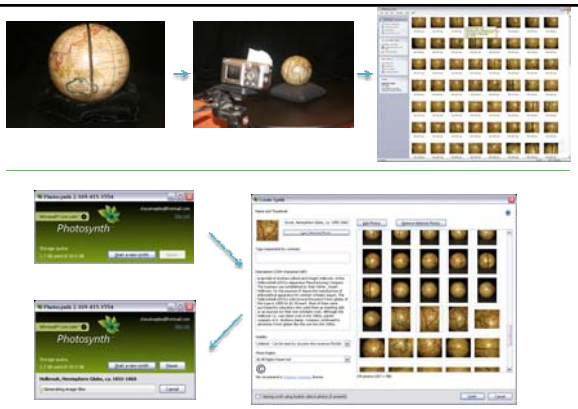
- What is Photosynth?
  - Photosynth is a web-based service that takes unstructured collections of photographs (of a single object or place) and computes:
    - the viewpoint from which each of the photos was taken
    - A sparse 3D model of the scene
    - image-to-model correspondences

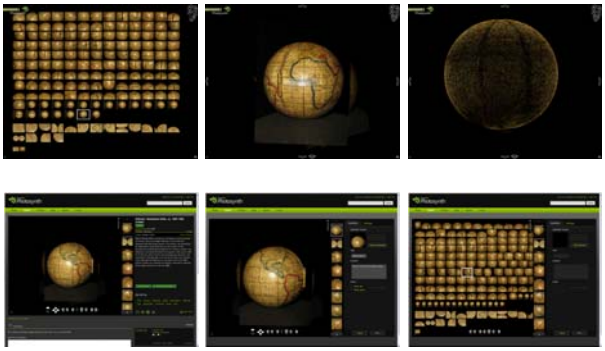
- Photosyth uses the Seadragon technology
- CBIR(Content-based Image Retrieval) - SIFT
- Arranges images in 3D space
- Uses extremely complicated algorithms to analyze camera angle, resolution, color, lighting...etc.

http://photosynth.net/



Demo





Demo

Alternatives to Photosynth?

- Bundler: Structure from Motion for Unordered Image Collections

- Bundler is an Open Source version of 'Photo Tourism' package that was the original project upon which Microsoft Photosynth is based.

- http://phototour.cs.washington.edu/bundler/

Q & A

講義內容 http://vcenter.blogspot.com/