LTU technologies

Image mining technologies and industrial challenges

Sébastien GILLES, Ph.D.
Chief scientist & co-founder
sg@ltutech.com
www.LTUtech.com

Context

- LTU has successfully deployed image mining softwares in very demanding industrial environments
  - Large data volumes, high throughput
  - Clients use the product and build value on it
  - Mission-critical tasks
  - Requirements: security, availability, failover, fast response time…and of course quality
- There are complex issues that need to be solved:
  - Adapt rapidly to changing conditions
    - Market
    - Economical, social environment
    - Technology
  - Design a generic and modular technology for multiple reuse
    - Standalone application
    - OEM
    - Integration

LTU corporate background

- Founded in 1999, LTU Technologies is a software company focused on image mining technologies.
  - Capitalizing on 10 years of high-profile research of the founders at MIT Media Lab, Oxford, INRIA
  - 20 employees, headquartered in Paris, office in Washington D.C.
- Market verticals
  - Law enforcement: World-wide deployment at top-level intelligence divisions (incl. FBI).
    - *Child exploitation, Stolen Art, Counterfeiting, ID theft.*
  - Industrial property: Patent offices, IP-protection companies run LTU.
    - *Trademark search, Brand protection, Counterfeiting.*
  - Asset management: e-Commerce, publishing companies
    - *Intra/Extra/Inter-net integration to Digital Asset Management softwares.*

LTU DNA vs. MD-5 or SHA-1

DNA

MD-5 or SHA-1

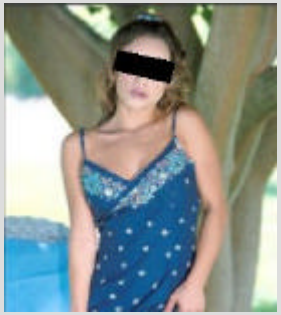


photo11.jpg

Image Database

100%

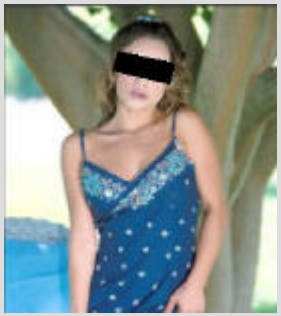
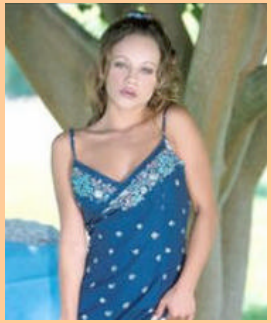


photo11.jpg

"Duplicates"

98%

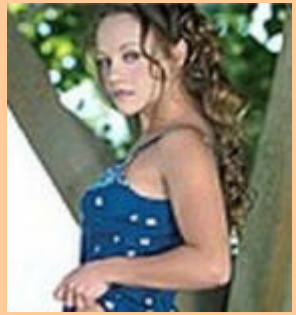
97%



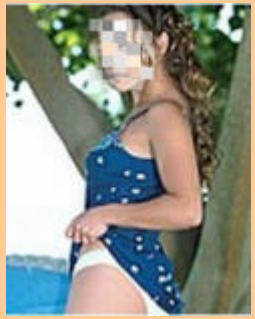
Young_girl.bmp

"Clones"

85%



80%



"Similar" images

Investigating with Image-Seeker

Image seizure
(HDD or Internet)

On the "ground"

Case: Barney
Victim? Abuse?
Date: 6/2/2004

Detection &
upload of seized
images

Local, Na
or Int'l
agencie

Image-Seeker™

Image
Database
+
Case
information

Victim Identification
Link to other cases

AO763456 XW787346 Anne Smith

Evidence of Abuse
Retrieval of series of images

AB763235 KW787386 KX9826563

Validation

Action

A Layered Software Architecture

HTTP

Image-Seeker HTML GUI
Apache PHP server

PHP plugin API

PHP Plugins

Java Clients

Windows Clients

SDK layer (php/java/perl/C/…)

XML/HTTP

Image-Seeker HTTP API

LTU Image-Seeker platform
Distributable and redundant (CORBA)
Modular architecture (polymorphic java components)

LTU Components Java API

LTU algorithmic Components
Image Processing, DNA computation, Data Retrieval, Automatic Keywording

Storage/Search Components
Textual Search, Data Storage (databases/files)

Client-specific system built by LTU or partner

-Verity
-Investigation soft.
-MAM software

-.net web service
-GUI
-3rd party system

-Oracle
-PostgreSQL
-Verity
-…

Industrial challenges for DB

- General system design considerations
  - Multimedia indexing require CPU-intensive processes
  - Forget about DB server running MM indexing process
    - Oracle's plug-ins useless for large-scale multimedia indexing
  - N processing nodes and M storages nodes
- Ensure a true « operational » scalability
  - Distributed, clustered architecture (failover)
  - Bottleneck: maintenance
    - Re-index 1M images while maintaining QoS ?
    - Synchronization of N image repositories and data warehouses
- Adapt data models to multimedia data
  - Increased complexity with video, images, web pages, etc.
  - Issue: models are task-dependent, due to performance issues.
- Increase performance, reduce response-time
  - DBMS remain slow to access data: need for memory caching
  - Fast Nearest-Neighbour search in high dimensions and large volumes

Industrial challenges for Computer Vision

- Real-life images are not « clean »
  - Highly variable acquisition conditions (uncontrolled lighting, etc.)
  - Multiple imaging defects (focus, over/under-exposition, etc.)
  - Collection-dependent artefacts
- Client requirements vary a lot
  - Heterogeneous image types (pictures, drawings, logos, etc.)
  - Highly variable definition of « what matters in an image »
  - Global vs. Local analysis
- Performance/Quality tradeoffs
  - DNA extraction+classification to be performed in near real-time
  - Search to be performed in near real-time
  - If asked, clients tend to favour quality vs. performance (insight: Moore law)

Conclusions

- Industry offers many practical, technical challenges
- Research solves many theoretical problems
- A gap remains, that needs to be bridged: a full client solution generally involves several technologies
  - NLP, Image Analysis, AI, etc. (e.g: trademark search)
- This means 3 main challenges !
  - Integrating those technologies is challenge#1
  - Addressing real-life data, scenarios and usages is challenge#2
  - Optimizing large-scale, complex systems is challenge#3
- Issues:
  - Difficulty of obtaining large volumes real-life data for academics
  - Software re-use too rare in academics
- Ideas:
  - Validate research algorithms on professional platforms… (LTU?)
  - Develop common benchmarking efforts on real data