CLARITY
clarity-centre.org

So what can we actually do with content-based video retrieval ?

Alan F. Smeaton

CLARITY: Centre for Sensor Web Technologies
Dublin City University
Alan.Smeaton@DCU.ie

UNIVERSITY COLLEGE DUBLIN • DUBLIN CITY UNIVERSITY • TYNDALL NATIONAL INSTITUTE

Issues with video

CLARITY
clarity-centre.org

Many of the issues associated with video in digital form are solved !
Capture, formatting, compression, storage, transmission, rendering on fixed and mobile

Outstanding challenges are in managing video content
..means analysis, indexing, summarising, browsing and searching

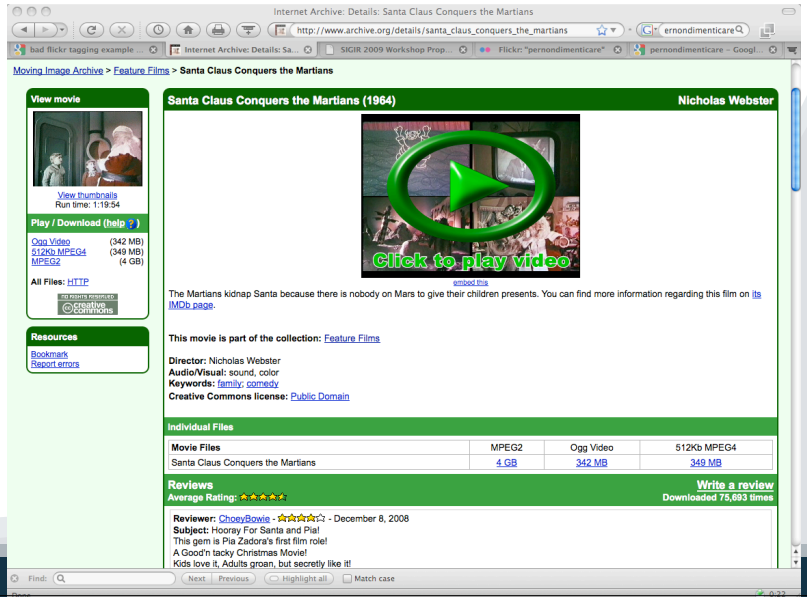
Managing video is mostly done with metadata
… title, date, actor, producer, genre, running time, format, reviews, ratings, etc. … and **user-generated tags** (UGC).

Some of these are coupled with keyframe / storyboard previews

3 examples ….

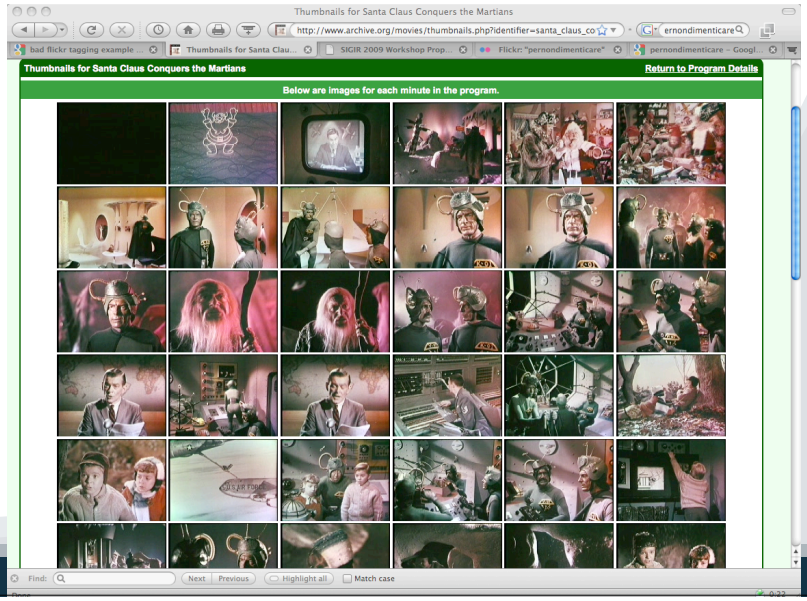
Internet Movie Archive

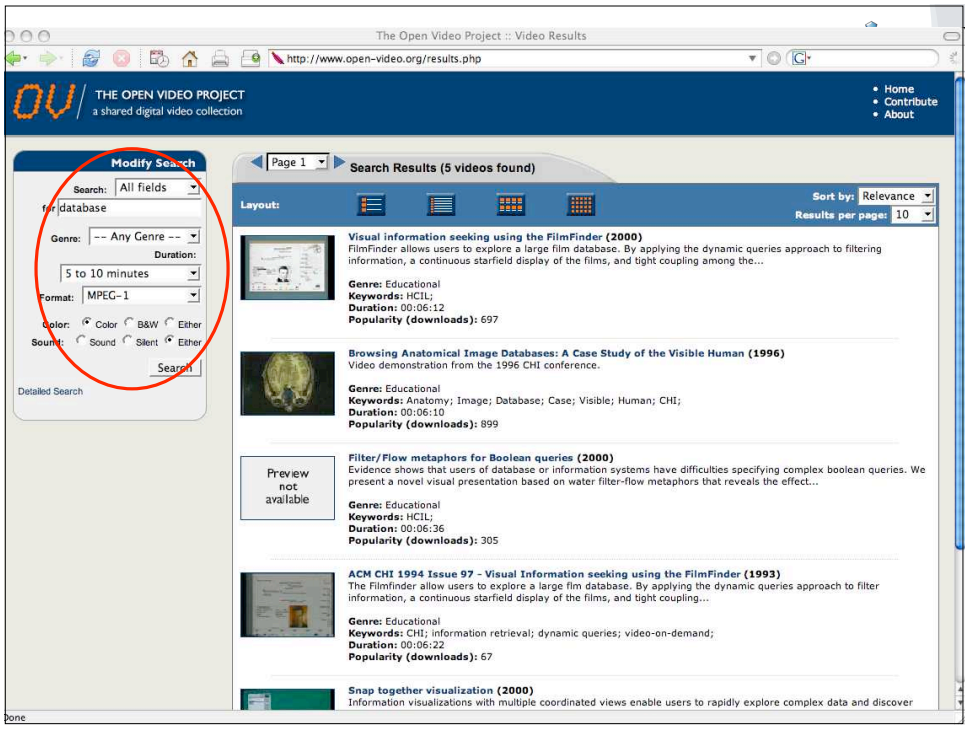
CLARITY
clarity-centre.org

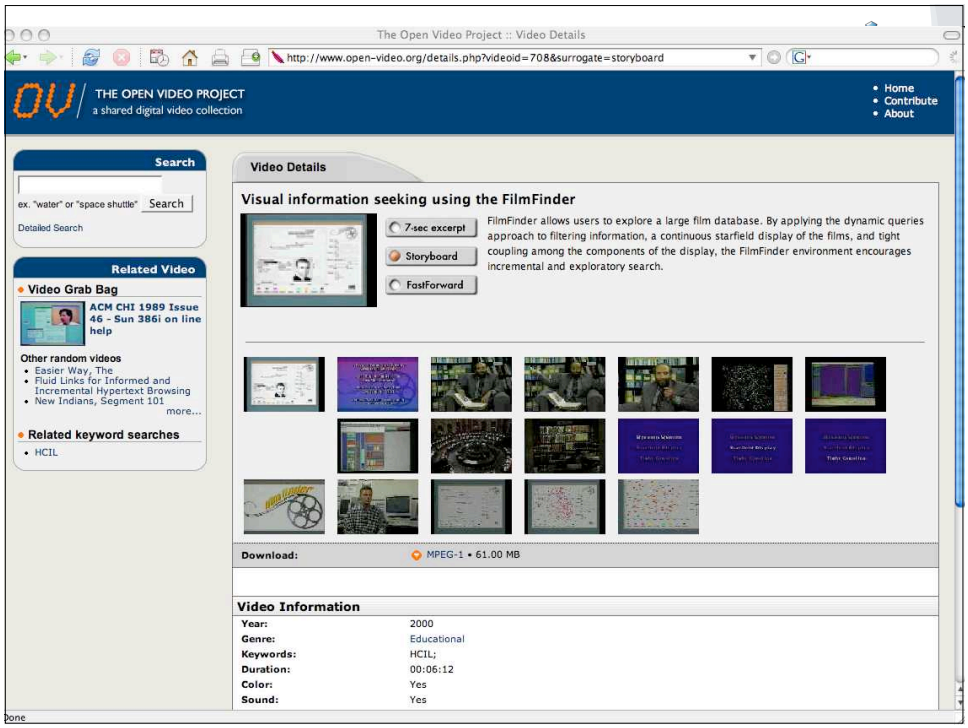


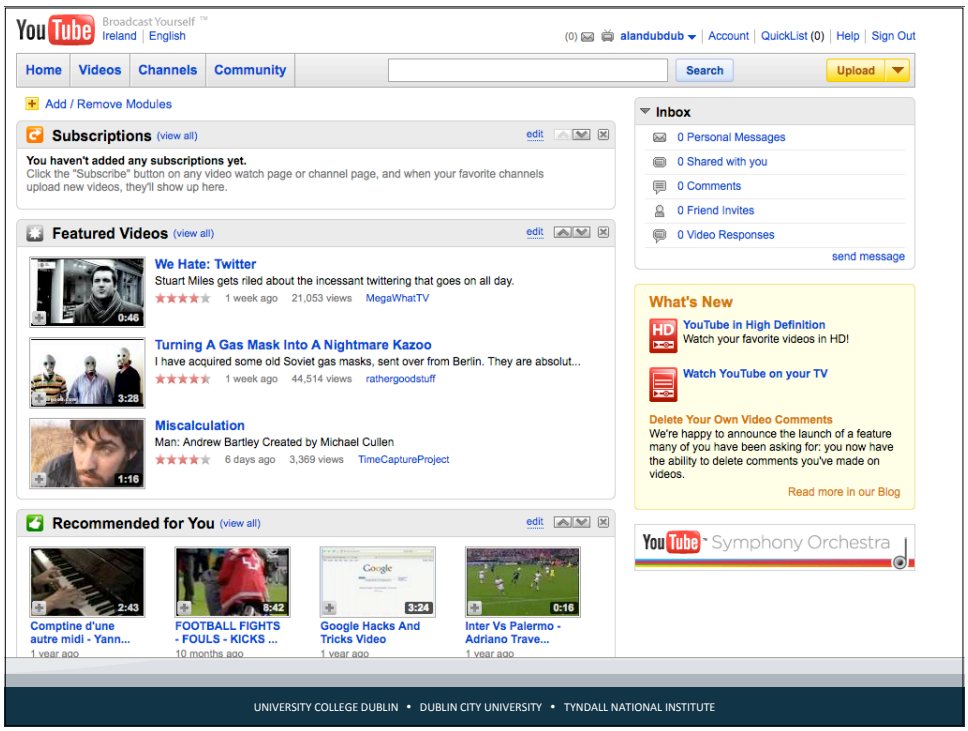
Internet Movie Archive

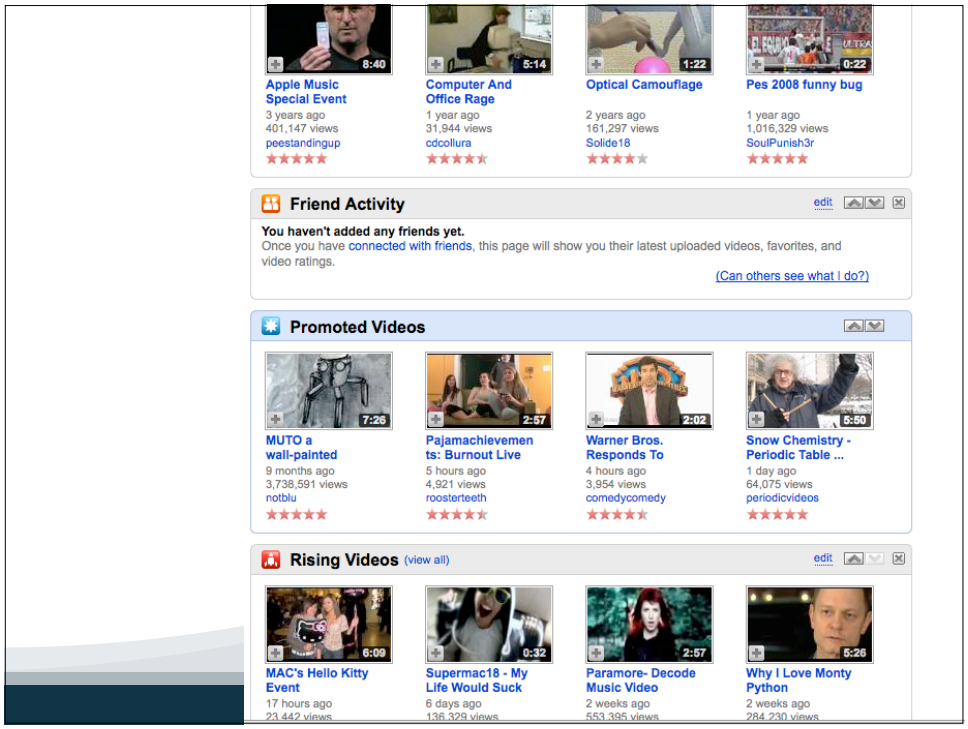
CLARITY
clarity-centre.org

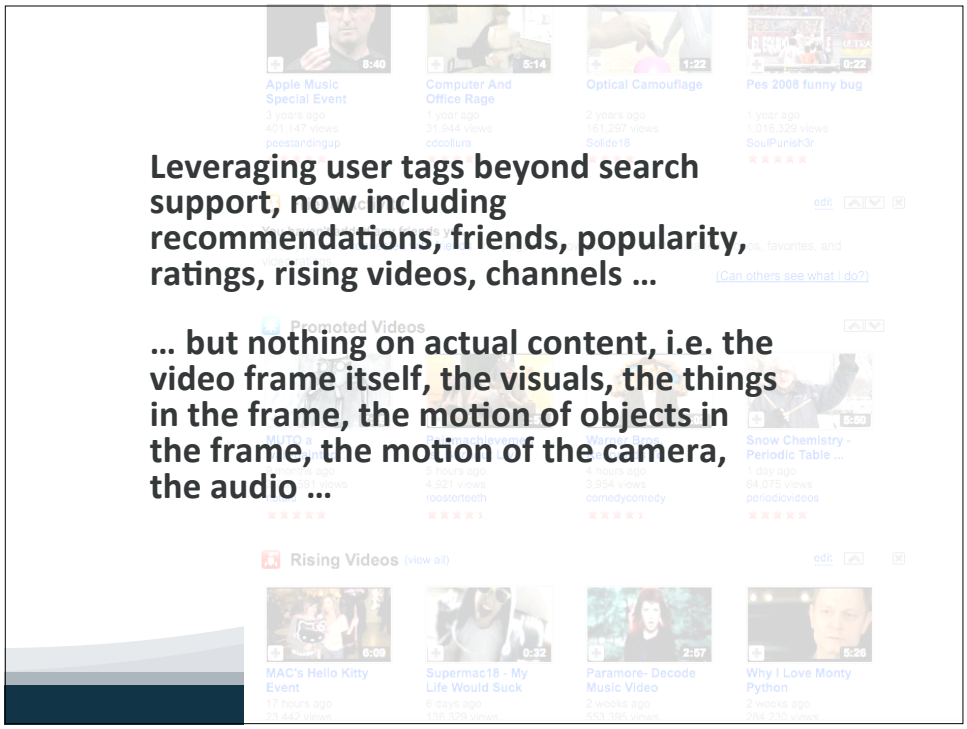





The Open Video Project :: Video Details
http://www.open-video.org/details.php?videoid=708&surrogate=storyboard
THE OPEN VIDEO PROJECT
a shared digital video collection
Home
Contribute
About
Search
ex. "water" or "space shuttle"
Search
Detailed Search
Related Video
Video Grab Bag
ACM CHI 1989 Issue 46 - Sun 386i on line help
Other random videos
Easier Way, The
Fluid Links for Informed and Incremental Hypertext Browsing
New Indians, Segment 101
more...
Related keyword searches
HCIL
Video Details
Visual information seeking using the FilmFinder
7-sec excerpt
Storyboard
FastForward
FilmFinder allows users to explore a large film database. By applying the dynamic queries approach to filtering information, a continuous starfield display of the films, and tight coupling among the components of the display, the FilmFinder environment encourages incremental and exploratory search.
Download:
MPEG-1 • 61.00 MB
Video Information
Year: 2000
Genre: Educational
Keywords: HCIL;
Duration: 00:06:12
Color: Yes
Sound: Yes
Done


YouTube
Broadcast Yourself
Ireland | English
(0) alandubdub | Account | QuickList (0) | Help | Sign Out
Home
Videos
Channels
Community
Search
Upload
Add / Remove Modules
Subscriptions (view all)
edit
You haven't added any subscriptions yet.
Click the "Subscribe" button on any video watch page or channel page, and when your favorite channels upload new videos, they'll show up here.
Featured Videos (view all)
We Hate: Twitter
Stuart Miles gets riled about the incessant twittering that goes on all day.
1 week ago 21,053 views MegaWhatTV
Turning A Gas Mask Into A Nightmare Kazoo
I have acquired some old Soviet gas masks, sent over from Berlin. They are absolut...
1 week ago 44,514 views rathergoodstuff
Miscalculation
Man: Andrew Bartley Created by Michael Cullen
6 days ago 3,369 views TimeCaptureProject
Recommended for You (view all)
Comptine d'une autre midi - Yann...
1 year ago
FOOTBALL FIGHTS - FOULS - KICKS ...
10 months ago
Google Hacks And Tricks Video
1 year ago
Inter Vs Palermo - Adriano Trave...
1 year ago
Inbox
0 Personal Messages
0 Shared with you
0 Comments
0 Friend Invites
0 Video Responses
send message
What's New
YouTube in High Definition
Watch your favorite videos in HD!
Watch YouTube on your TV
Delete Your Own Video Comments
We're happy to announce the launch of a feature many of you have been asking for: you now have the ability to delete comments you've made on videos.
Read more in our Blog
YouTube Symphony Orchestra
UNIVERSITY COLLEGE DUBLIN • DUBLIN CITY UNIVERSITY • TYNDALL NATIONAL INSTITUTE





Leveraging user tags beyond search support, now including recommendations, friends, popularity, ratings, rising videos, channels …

… but nothing on actual content, i.e. the video frame itself, the visuals, the things in the frame, the motion of objects in the frame, the motion of the camera, the audio …

Content based video navigation

CLARITY
clarity-centre.org

… is what I'm interested in. There are 4 approaches:

- **Use text from speech - ASR/CC/in-video OCR**
- **Match keyframes vs. query images**
- **Use semantic video features**
- **Use video/image objects as queries**

… and I could happily show examples of our systems in each class .. but lets not do that .. lets look at how video systems can be benchmarked, lets look at TRECVid;

TRECVid goals and strategy

CLARITY
clarity-centre.org

Promote progress in content-based analysis, detection, retrieval on large amounts of digital video

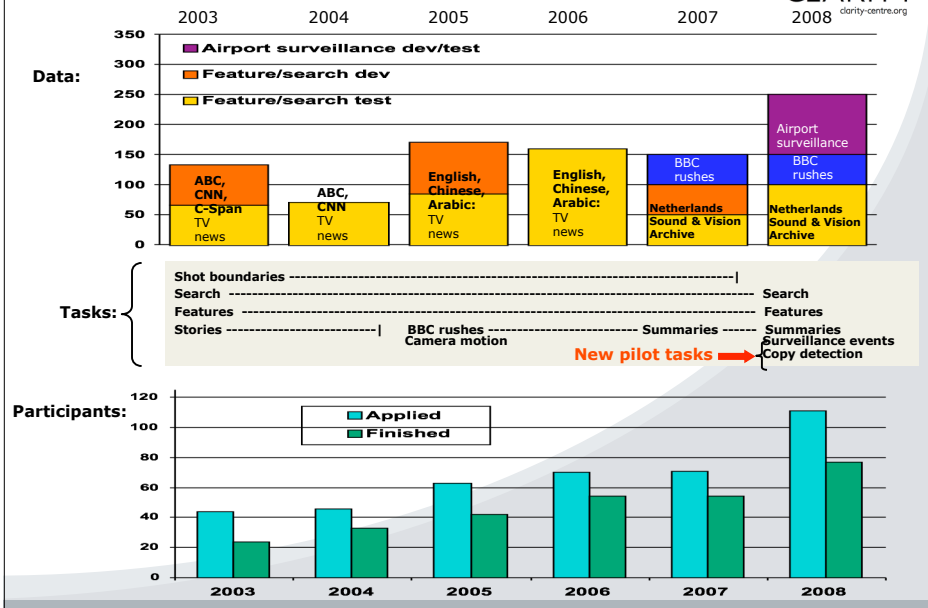
Measure systems against human abilities

Focus on relatively high-level functionality – near that of an end-user application like interactive search

Supplement with focus on supporting and related automatic components:
Automatic search, high-level feature detection, shot bound detection, content-based copy detection, event detection

Do all this in a hugely collaborative and supportive framework

Evolution: data, tasks, participants, ...



2008: Details

Data:

–200 hrs - Netherlands Institute for Sound and Vision (S&V)
–40 hrs - BBC rushes
–100 hrs of airport surveillance data - UK Home Office

5 evaluated tasks

–Content-based copy detection – 2010 video queries,…
–High-level feature extraction - 20 features
–Search (automatic, manually-assisted, interactive) - 48 topics
–Video summarization
–Event detection on airport surveillance video (5 cameras * 2 hours * 10 days)

TV2008 Finishers

Athens Information Technology
Asahikasei Co.
AT&T Labs - Research
Beckman Institute
Bilkent University
University of Bradford
Beijing Jiaotong University
Brno University of Technology
Beijing University of Posts and Telecommunications
Carnegie Mellon University
Columbia University
Computer Research Institute of Montreal
COST292 Team (Delft Univ.)
cs24_kobe (Kobe Univ.)
Dublin City University
ETIS Laboratory
EURECOM
Florida International Univ.
Fudan University
FX Palo Alto Laboratory
IBM T. J. Watson Research Center
INRIA-LEAR
INRIA-IMIA
IntuVision, Inc.
Ipan_uoi (University of Ioannina)
IRIM
ISM (The Institute of Statistical Mathematics)
Istanbul Technical University
IUPR-DFKI
JOANNEUM RESEARCH Forschungsgesellschaft mbH
KB Video Retrieval
K-Space
LIG (Laboratoire d'Informatique de Grenoble)
Laboratoire LIRIS (LYON)
University of Twente and CWI
LSIS_GLOT(CNRS LSIS)
Marburg
Chinese Academy of Sciences (MCG-ICT-CAS)
Mediamill (Univ. of Amsterdam)
MESH
MMIS (Open Univ.)
Microsoft Research Asia
NHKSTRL
National Institute of Informatics
National University of Singapore
National Taiwan University

TV2008 Finishers

NTT Cyber Solutions Laboratories
Orange Labs - France Telecom Group
Osaka University
Oxford Univ.
PKU-ICST (Peking Univ.)
PicSom (Helsinki University of Technology)
Queen Mary University of London
Queensland University of Technology
REGIM
Shanghai Jiao Tong University (SJTU)
SP-UC3M (Universidad Carlos III de Madrid)
The Hong Kong Polytechnic University
Tsinghua University - Intel China Research Center
Tsinghua University
TNO-ICT
Toshiba Corporation
Tokyo Institute of Technology
University of Alabama
University of Electro-Communications
University of Glasgow
University of Karlsruhe (TH)
University of Ottawa - SITE
University of Sheffield
University of Southern California
Universidad Rey Juan Carlos
Universidad Autonoma de Madrid
Universite Pierre et Marie Curie - LIP6
VIREO (City University of Hong Kong)
vision@ucf (University of Central Florida)
VITALAS (CERTH-ITI (GR), CWI(NL), U.Sunderland (UK))
XJTU (Xi'an Jiaotong University)

UNIVERSITY COLLEGE DUBLIN • DUBLIN CITY UNIVERSITY • TYNDALL NATIONAL INSTITUTE

Additional resources and contributions

City University of Hong Kong, the Laboratoire d'Informatique de Grenoble, and the University of Iowa helped out in the **distribution of video data** by mirroring the them online.

Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin provided the **master shot reference**

Roeland Ordelman and Marijn Huijbregts at the University of Twente donated the output of their **automatic speech recognition** system run on the Sound and Vision data

Christof Monz of Queen Mary, University London, who contributed **machine translation (Dutch to English)** for the Sound and Vision video.

INRIA's Nozha Boujemaa, Alexis Joly, and Julien.Law-to led the design of the **copy detection task**, in particular regarding the definitions of the video transformations. They provided an independent person, Laurent Joyeux, who created original queries and applied the 10 video transformations in a process blind to the ground truth.

Dan Ellis at Columbia University devised and applied the audio transformations to **produce the audio-only queries** for copy detection.

Additional resources and contributions

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble) organized a **collaborative annotation of 2008 development data** for 20 features. 40 groups contributed a total of 1.2 million image x concept annotations.

The Multimedia Content Group at the Chinese Academy of Sciences provided full **annotation of test features** for 2008 training data including location rectangles for object features.

Columbia University and the City University of Hong Kong contributed **detection scores for the 2008 data**: CU-VIREO374.

The University of Amsterdam provided 2 benchmarks for assessing **mappings of topics to concepts** for video retrieval.

Phil Kelly at Dublin City University (DCU) assisted with the **assessment of the rushes** summaries.

Carnegie Mellon University created a **baseline summarization** run to help put the summarization results in context.

UNIVERSITY COLLEGE DUBLIN • DUBLIN CITY UNIVERSITY • TYNDALL NATIONAL INSTITUTE

Tasks …

Lets briefly look at feature detection, and interactive search … but first SBD …

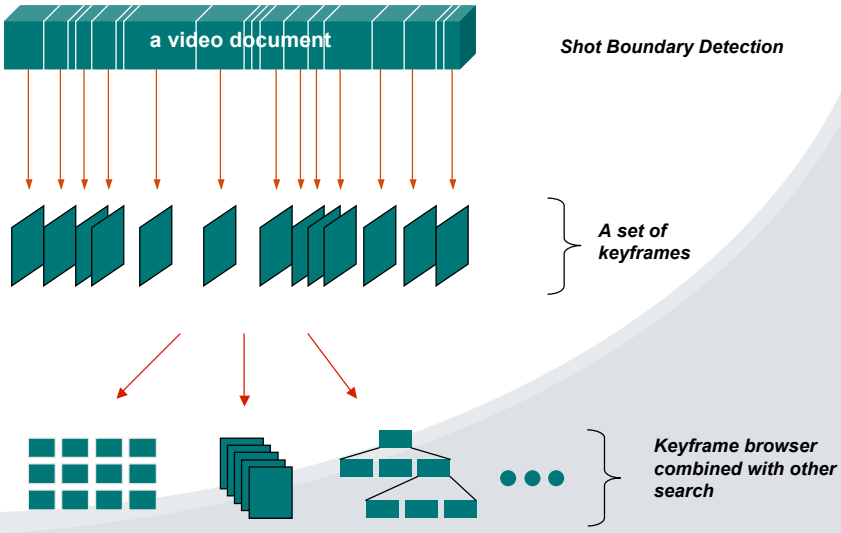
Automatic structuring of video is necessary to make progress in managing, a.k.a. shot bound detection;

A shot in video information is a sequence of continuous images (frames) from a single camera;

SBD was run for several years, manual annotation of ground truth, covering hard cuts and gradual transitions;

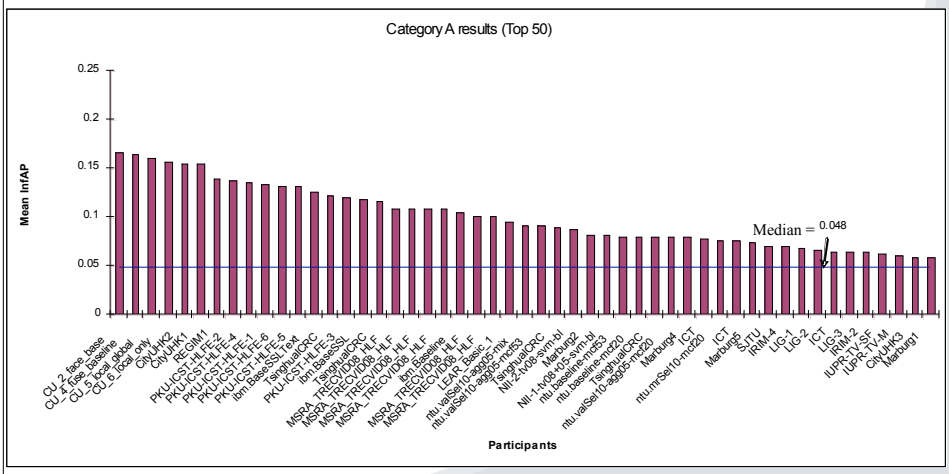
The task of SBD or automatic video segmentation is to segment video into its constituent shots … it's a solved problem … 95% P/R for hard cuts, 70% P/R for GTs, 1%-2% real time on standard desktops, not even using GPUs;

Automatic structuring of video

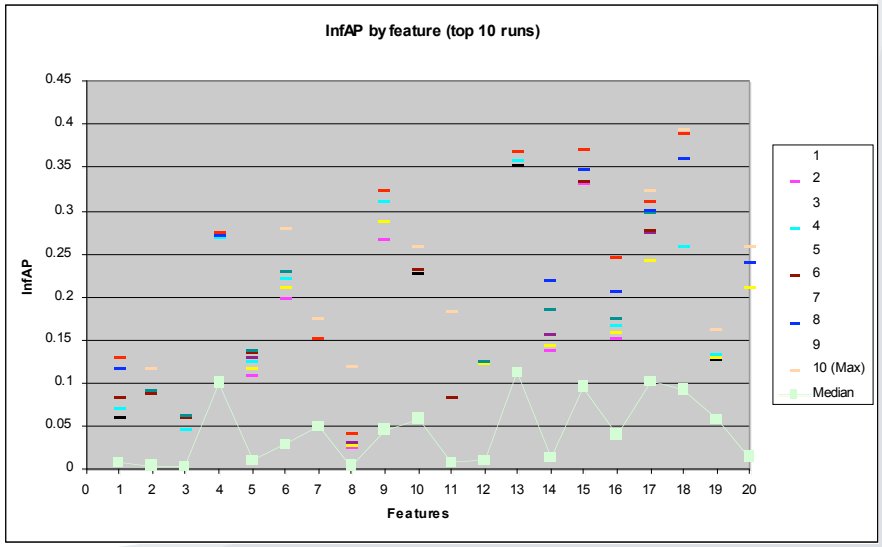


20 LSCOM features evaluated

1 Classroom
2 Bridge
3 Emergency_Vehicle
4 Dog
5 Kitchen
6 Airplane_flying
7 Two people
8 Bus
9 Driver
10 Cityscape
11 Harbor
12 Telephone
13 Street
14 Demonstration_Or_Protest
15 Hand
16 Mountain
17 Nighttime
18 Boat_ship
19 Flower
20 Singing



1 Classroom 2 Bridge 3 Emergency_Vehicle 4 Dog 5 Kitchen 6 Airplane_flying 7 Two people 8 Bus
9 Driver 10 Cityscape 11 Harbor 12 Telephone 13 Street 14 Demonstration_Or_Protest 15 Hand
16 Mountain 17 Nighttime 18 Boat_ship 19 Flower 20 Singing

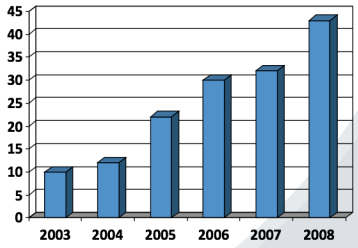


General observations (1)

CLARITY
clarity-centre.org

Participation is still increasing

Accepted as an important building block for search



General observations (2)

Hardly any feature specific approaches

Large variety in classifier architectures and choices of feature representations

Hardware: usually a single, cpu, but several medium and larger clusters

Nr of classifiers used for fusion ranges between 1 and >1160

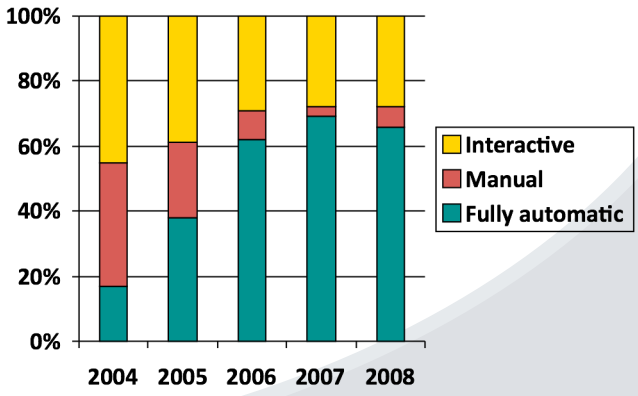
Testing times vary between 10m and 150h per feature.

Approx. 30% of the runs do some form of temporal analysis

Approx 50% of the runs use salient/SIFT points

Remember - these are features PER SHOT - not per scene !

TRECVid Search … trends stable



24 Topics (for all systems)

Find shots of a person opening a door
Find shots of 3 or fewer people sitting at a table
Find shots of one or more people with one or more horses
Find shots of a road taken from a moving vehicle, looking to the side
Find shots of a bridge
Find shots of one or more people with mostly trees and plants in the background; no road or building can be seen
Find shots of a person's face filling more than half of the frame area
Find shots of one or more pieces of paper, each with writing, typing, or printing it, filling more than half of the frame area
Find shots of one or more people where a body of water can be seen
Find shots of one or more vehicles passing the camera
Find shots of a map
Find shots of one or more people, each walking into a building
Find shots of one or more black and white photographs, filling more than half of the frame area
Find shots of a vehicle moving away from the camera
Find shots of a person on the street, talking to the camera
Find shots of waves breaking onto rocks
Find shots of a woman talking to the camera in an interview located indoors - no other people visible
Find shots of a person pushing a child in a stroller or baby carriage
Find shots of one or more people standing, walking, or playing with one or more children
Find shots of one or more people with one or more books
Find shots of food and or drinks on a table
Find shots of one or more people, each in the process of sitting down in a chair
Find shots of one or more people, each looking into a microscope
Find shots of a vehicle approaching the camera

Some approaches

University of Amsterdam (MediaMill)
- Optimal query mode (speech, detector, or example-based search) prediction by topic

Chinese Academy of Sciences (MCG-ICT-CAS)
- Distribution based concept selection method
- SIFT visual-keywords feature in low dimensional LDA semantic space
- Re-ranking based on the motion and face
- Dynamic fusion based on the Smoothed Similarity Cluster

K-Space
- Large multi-site interactive search experiment

FX Palo Alto
- Using program-based clustering to enhance search
- Collaborative search

Participant approaches

Brno University of Technology
Automatic runs using ASR and HLFs

Columbia University
Interactive runs using CuZero browser exploring novice vs. expert, query formulation vs. full browser experience, story-based expansion

Cost292
A mere 36 co-authors (vs. 40 in K-Space ;-) so a large multi-site group effort
Text, visual and HLF interactive search plus audio filtering, term recommendation and relevance feedback

cs24_kobe (Kobe Univ.)
Use multiple examples per topic, and rough set theory to "conceptualise" the topic, leading to interactive retrieval

Dublin City University
Automatic runs, focus on query time weights for fusion from different retrieval experts

Participant approaches

Fudan University
Automatic runs to explore fusions of text, visual and HLF-based retrieval

IBM
Interactive runs varying the number of HLFs available and the impact of near-duplicate detection and shot clustering

KBVR (David Etter)
Using text and image features and exploring augmentation with knowledge from Wikipedia and form image clusters

U. Twente / CWI (Lowlands Team)
Automatic runs varying the set of concepts (M'Mill 101 and VIREO 374) and also Wikipedia articles for text expansion

MMIS (Open U, moved from Imperial College)
Another multi-site group, first timers. Submitted text-only plus automatic run based on MPEG-7 visual features

Participant approaches

Microsoft Research Asia
Automatic runs with text and visual baselines, query-independent learning, and various reranking methods

National Institute of Informatics, Japan
Automatic runs with concept suggestion based on text query vs text descriptions of LSCOM 374 HLFs

National University of Singapore

National Taiwan University

Oxford University
Same system as 2007 (useful!), visual-only interactive search
System included additional external images from Google search and detection of near-duplicates, upper body and face

Participant approaches

Helsinki University of Technology (PicSOM)
Automatic runs focusing on text+HLFs only; when HLFs not possible, only then do visual based search; also included face detection and motion features

REGIM (ENIS, Tunisia)
Interactive search, fusion of text & HLFs plus detection of faces, vehicle, on-screen text and 1+ people

Shanghai Jiao Tong University, Shanghai
Automatic search using text, 20x HLFs and QBE using colour moments

SP-UC3M (Universidad Carlos III de Madrid)

Tsinghua University / Intel China
Automatic runs, use rich image features to build a SVM for each topic; also use user tags on Flickr images to locate extra images for example-based search; fuse all combinations

Participant approaches

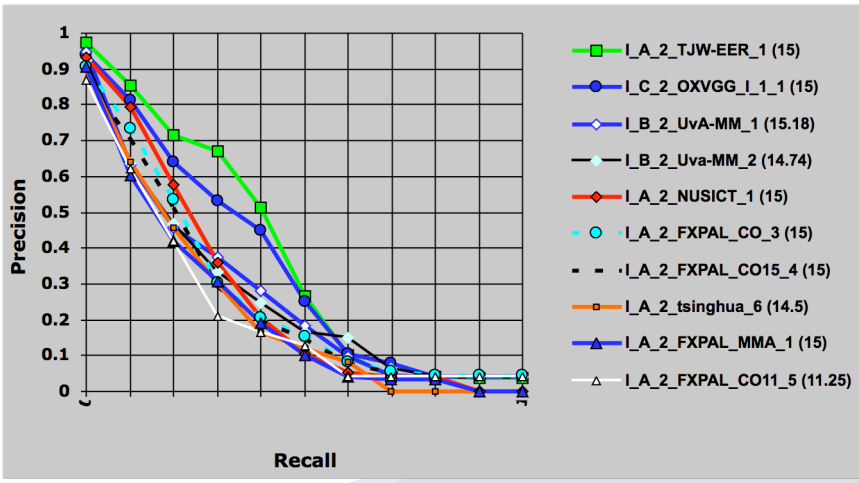
University of Alabama (with UNC)
Manual & interactive, text plus QBE using image features

University of Glasgow
Automatic runs using text, MPEG-7 visual features, HLFs and image classification using SVMs, and an interactive run which clusters/groups similar results

VIREO - City University of Hong Kong
Automatic search on HLFs only considering fusion of detectors using concept semantics, co-occurrence, diversity, and detector robustness

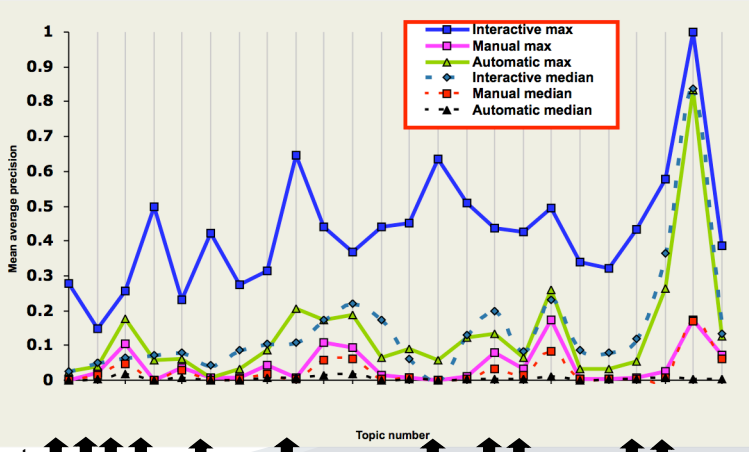
VITALAS (Thessaloniki, ITI Crete, CWI & Twente)
Focus on concept retrieval, combine text and HLFs merge (text) concept descriptors proportional to Prob of occurrence

'07 Interactive runs - top 10 MAP (of 33)



Another view: in highest scoring run, on average 8 of the top 10 shots returned contained the desired video

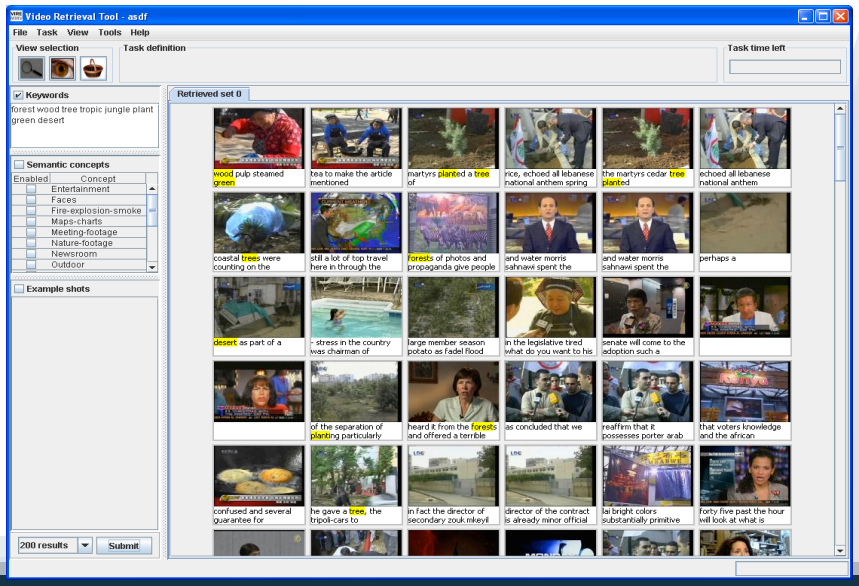
Average precision by topic (07)



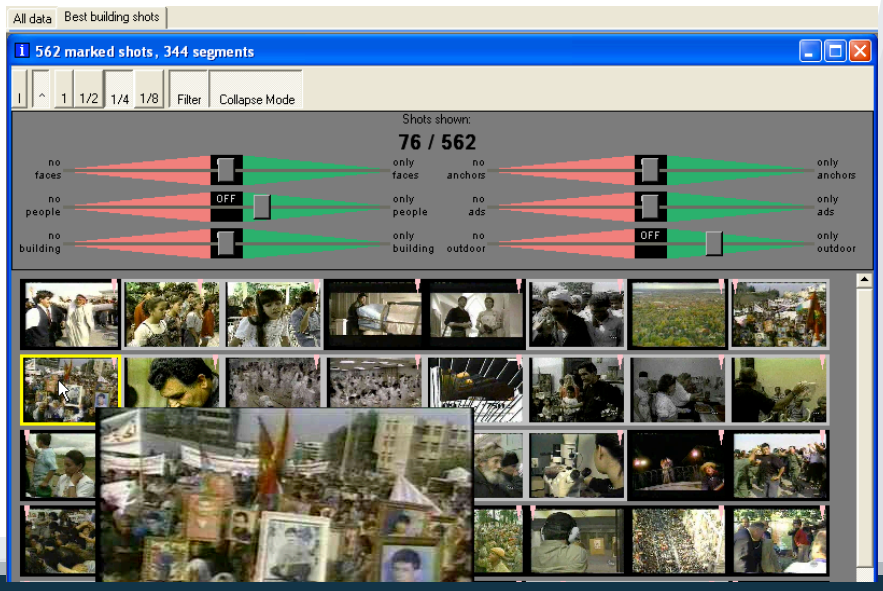
Events

What do systems look like ?

U Oulu


Video Retrieval Tool - asdf
File Task View Tools Help
View selection
Task definition
Task time left
Keywords
forest wood tree tropic jungle plant green desert
Semantic concepts
Enabled Concept
Entertainment
Faces
Fire-explosion-smoke
Maps-charts
Meeting-footage
Nature-footage
Newsroom
Outdoor
Example shots
200 results
Submit
Retrieved set 0

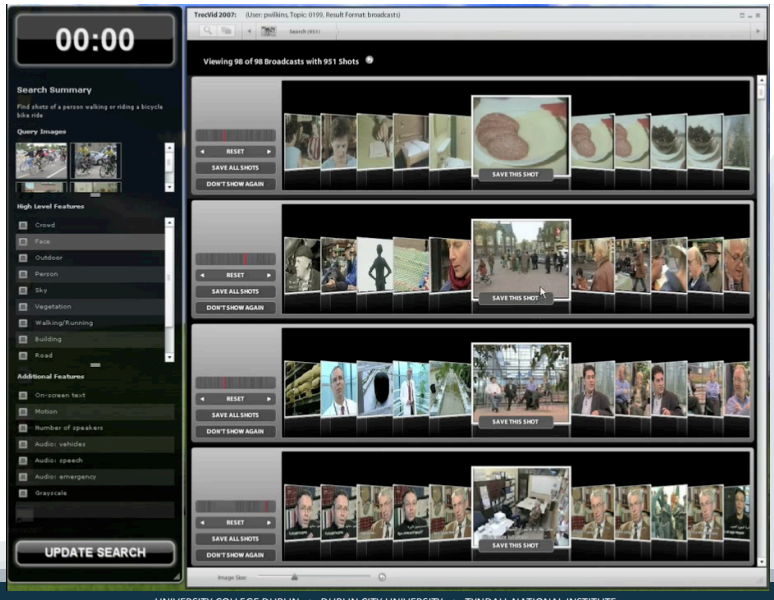
CMU


All data
Best building shots
562 marked shots, 344 segments
1
1/2
1/4
1/8
Filter
Collapse Mode
Shots shown:
76 / 562
no faces
only faces
no anchors
only anchors
no people
OFF
only people
no ads
only ads
no building
only building
no outdoor
OFF
only outdoor

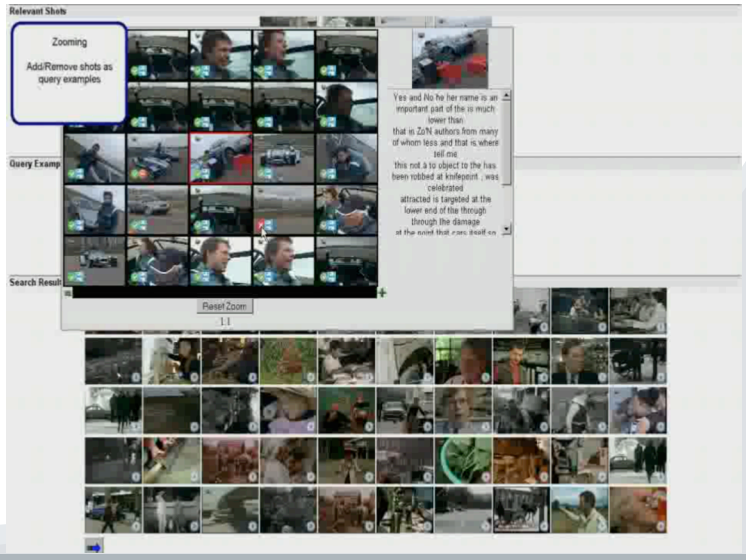
DCU


00:00
Search Summary
Query Images
High Level Features
Crowd
Face
Outdoor
Person
Sky
Vegetation
Walking/Running
Building
Additional Features
On-screen text
Motion
Number of speakers
Audio: vehicles
Audio: speech
Audio: emergency
Grayscale
UPDATE SEARCH
TrecVid 2007:
Viewing 949 of 949 Results
REMOVE
SAVE
BROADCAST

DCU


00:00
Search Summary
Query Images
High Level Features
Crowd
Face
Outdoor
Person
Sky
Vegetation
Walking/Running
Building
Road
Additional Features
On-screen text
Motion
Number of speakers
Audio: vehicles
Audio: speech
Audio: emergency
Grayscale
UPDATE SEARCH
TrecVid 2007:
Viewing 98 of 98 Broadcasts with 951 Shots
RESET
SAVE ALL SHOTS
DON'T SHOW AGAIN
SAVE THIS SHOT

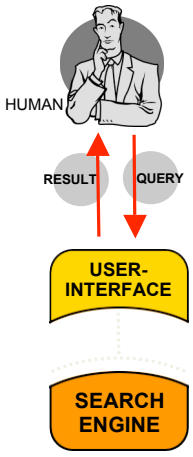
U Glasgow



Each year we showcase interactive TRECVid video search at the CIVR conference

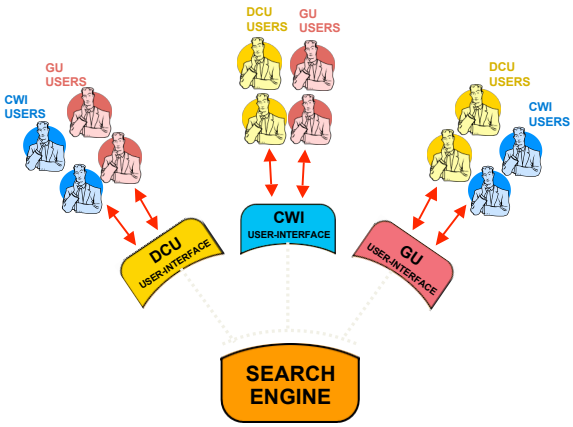
Called the VideOlympics …

What do participants do ?



K-Space participation 2008

What do participants do ?



K-Space participation 2008

UNIVERSITY COLLEGE DUBLIN • DUBLIN CITY UNIVERSITY • TYNDALL NATIONAL INSTITUTE

State of the art ?

CLARITY
clarity-centre.org

On small, closed video libraries, content based video search works well; with metadata and UGC it would be even better …

We're still only doing keyframe/image and not video (with motion of objects and cameras), and we're purposely not using metadata or tags or UGC;

We're still doing shot retrieval, not scene, or clip … that's the task;

Feature detection accuracy, scale-up to more features, relationships between features, move away from independent solo to ontology-based … need to progress this;

Combining features, keyframe match, text and objects in a natural and usable way … the learnability of the interface;

Dynamically adjusting retrieval to the query/video type;

Challenges …

CLARITY
clarity-centre.org

Not participation, not organisation, not tasks, not scientific rigour, not enthusiasm, not research topics … it's the data !

NIST cannot legally distribute data which is not 100% © cleared to do so .. LDC, S&V

We need help with data !

Consequence is that data too irrelevant ?

Too closed shop, not public enough ?
VideOlympics showcase, Summarisation workshop at ACM MM

Learnability of systems for non-experts ?
Most sites used expert searchers … recent paper showed searcher variability across sites to be a factor … VideOlympics '09 uses schoolchildren !

Too US DTO ?
See the list of contributors !