Video Recognition and Retrieval at the TRECVID Benchmark

Introduction

George Awad, Paul Over {retired} – National Institute of Standards & Technology
Alan Smeaton – Dublin City University
Cees Snoek, Arnold Smeulders – University of Amsterdam
Shin'ichi Satoh – National Institute of Informatics
Kazuya Ueki -- Meisei University; Waseda University

Points to be covered

- What is TRECVID
- TRECVID video datasets
- Popular tasks at TRECVID
- Important impacts
- Latest stats in 2017
- Observations and questions
- What's new in TRECVID 2019

What is TRECVID ?

- Workshop series (2001 – present) → http://trecvid.nist.gov
- Started as a track in the TREC (Text REtrieval Conference) evaluation benchmark.
- Became an independent evaluation benchmark since 2003.
- Focus: **content-based** video analysis, retrieval, detection, etc
- Provides data, tasks, and uniform, appropriate scoring procedures
- Aims for realistic system tasks and test collections:
  - unfiltered data
  - focus on relatively high-level functionality (e.g. interactive search)
  - measurement against human abilities
- Forum for the
  - exchange of research ideas and for
  - the discussion of research methodology – what works, what doesn't , and why

TRECVID Philosophy

- TRECVID is a modern example of the Cranfield tradition
  - Laboratory system evaluation based on test collections
- Focus on advancing the state of the art from evaluation results
  - TRECVID's primary aim is **not** competitive product benchmarking
  - experimental workshop: sometimes experiments fail!
- Laboratory experiments (vs. e.g., observational studies)
  - sacrifice operational realism and broad scope of conclusions
  - for control and information about causality – what works and why
  - results tend to be narrow, at best indicative, not final
  - evidence grows as approaches prove themselves repeatedly, as part of various systems, against various test data, over years

Drilling deeper in the search landscape

You want something to make you laugh

Voter looks for video of candidate X at recent town hall meeting

Your mother searches home videos for shots of daughter playing with family pet.

10-yr old looks for video of tigers for school report

Doctor searches echocardiogram videos for instances like example

Security personnel searches surveillance video archive for suspicious behavior

Student searches Web for new music video

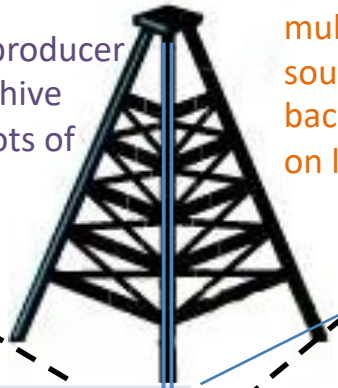
Fan searches for favorite TV show episode

Documentary producer searches TV archive for reusable shots of Berlin in 1920's

Intelligence analyst searches multilingual open source video for background info on location X

Searcher abilities, needs, preferences, history

Data types



Human-computer interaction	Human visual capabilities, expert vs novice, text/image/concept querying, visualization, …
Information retrieval	Indexing, query typing, concept selection, weighting, ranking, pos/neg relevance feedback, metadata, …
Machine vision	Segmentation, keypoints, SIFT, classifier fusion, face recognition, …
Machine learning	CNNs, DL, SVM, GMM, graphical models, boosting, …
Metrology …	Metrics, data, task definition, ground truth, significance, …

TRECVID search types

- TRECVID search has modeled a user looking for video shots
  - Of people, objects, locations, actions, events, activities
  - Not just information (e.g., video of X, not video of someone talking about X)
  - Independent of original intent, saliency, etc.
- In video of various sorts :
  - Multilingual broadcast news (Arabic, Chinese, English)
  - Dutch “edutainment”, cultural, news magazine, historical shows
  - Soap opera episodes
  - Internet videos (Social media, movies, programs)
  - Security cameras
- Video data diverse in :
  - Content, style, means of production, associated metadata, possible delivery platforms, etc
- Using queries containing:
  - text only –
  - text + image/video examples
  - image/video examples only
- In two modes:
  - fully automatic
  - human-in-the-loop search (Interactive , manually-assisted)

TRECVID Datasets overview

US TV news (`03/`04)

International TV news (`05/`06)

Dutch TV infotainment (`07/`08/`09)

Web video (since 2010)

Soap opera (since 2013)

CNN abc

LBC

CCTV NBC

BEELD EN GELUID

HAVIC

INTERNET ARCHIVE

Security cameras (since 2008)

Social media (since 2016)

Vine

How difficult can video search be?

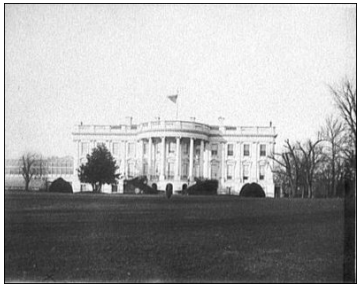
One image/video – many different (changing) views of content

Possible content keywords, tags:

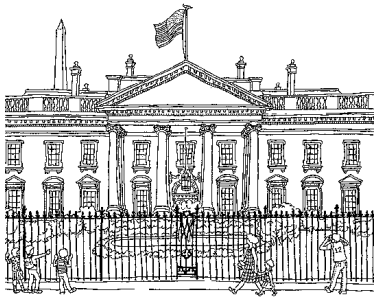
women
pigeons
plaza
buildings
outdoors
daytime
running
falling
clapping
….

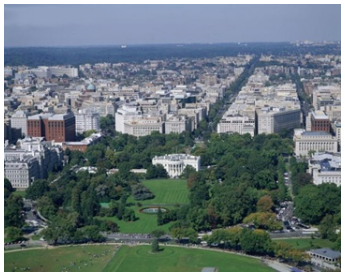
Creator's keywords: "stupid sister"

One person/thing/location – many different (changing) appearances

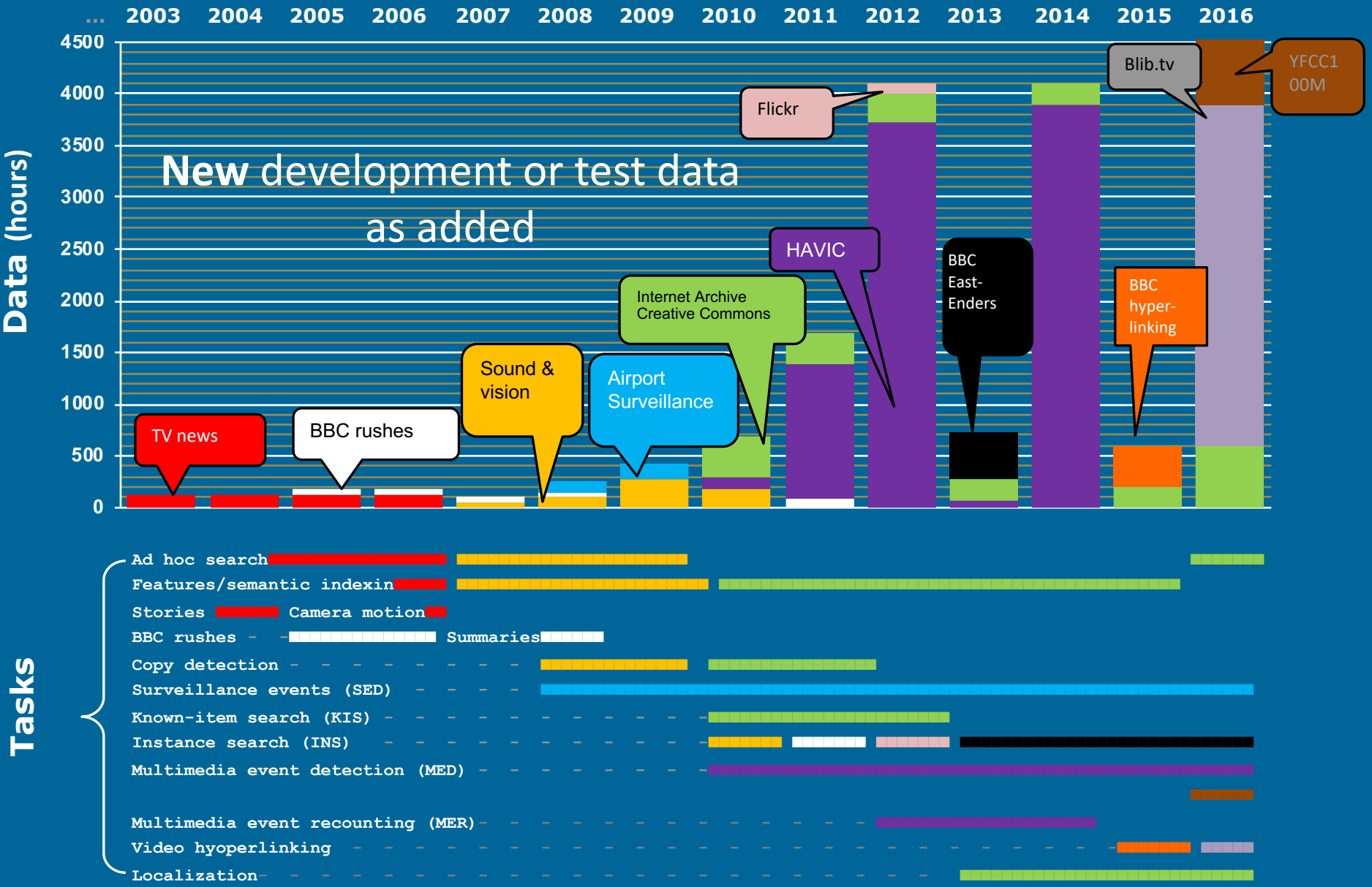





TRECVID Data/Tasks Evolution



Popular tracks/tasks at TRECVID (2001 – 2018)

Shot boundary detection (2001 – 2007)

- Shot boundary detection is a fundamental task in any kind of video content manipulation
- Task provided a good entry for groups who wish to “break into” video retrieval and TRECVID gradually
- **Task** : identify the shot boundaries with their location and type (cut or gradual) in the given video clip(s)

Video Search (aka Ad-hoc) (2001 – 2009, 2016 - present)

- The search task models a user who is looking for segments of video containing persons, objects, events, locations, etc. of interest
- Can contain specific named entities OR generic classes.
- Given the video search test collection, a statement of information need (topic), and the common shot boundary reference for the test collection, return a ranked list of at most 1000 common reference shots from the test collection, which best satisfy the need
- Topics/Queries can be text + optional image(s)/video(s) examples.

Panofsky/Shatford mode/facet matrix

**	Specific (Iconographic)	Generic (Pre-iconographic)	Abstract (Iconological)
Who	Individually named person, group, thing	Kind of person, thing	Mythical, fictitious being
What	Individually named event, action	Kind of event, action, condition	Emotion, abstraction
Where	Individually named geographical location	Kind of place, geographical, architectural	Place symbolized
When	Linear time: date or period	Cyclical time: season, time of day	Emotion, abstraction symbolized by time

** From Enser, Peter G. B. and Sandom, Chriss J. Retrieval of Archival Moving Imagery – CBIR Outside the Frame. CIVR2002. LNCS 2383 pp. 206-214.

Sample topics from TRECVID ad hoc search

- Find shots of a road taken from a moving vehicle through the front window.
- Find shots of a crowd of people, outdoors, filling more than half of the frame area.
- Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible.
- Find shots of a person talking on a telephone.
- Find shots of a close-up of a hand, writing, drawing, coloring, or painting.
- Find shots of exactly two people sitting at a table.
- Find shots of one or more people, each walking up one or more steps.
- Find shots of one or more dogs, walking, running, or jumping.
- Find shots of a person talking behind a microphone.
- Find shots of a **building entrance**.
- Find shots of **people shaking hands**.
- Find shots of a **microscope**.
- Find shots of a person pointing.
- Find shots of a person playing a piano.
- Find shots of a street scene at **night**.
- Find shots of printed, typed, or handwritten text, filling more than half of the frame area.
- Find shots of something burning with flames visible.
- Find shots of one or more people, each at a table or desk with a computer visible.
- Find shots of an airplane or helicopter on the ground, seen from outside.
- Find shots of one or more people, each sitting in a chair, talking.
- Find shots of one or more ships or boats, in the water.
- Find shots of a train in motion, seen from outside.
- Find shots with the camera zooming in on a person's face.
- Find shots of two more people, each singing and/or playing a musical instrument.

Some search results

Keyframes from top 20 clips returned by a system to query for "shots of person seated at computer "

High-Level Feature Extraction (aka Semantic Indexing) (2002 – 2015)

- Task goal was to encourage research in concept detection.
- Secondary goal was to allow usage of system outputs in the video search task.
- Task: Given the test collection, master shot reference, and concept definitions, return for each concept a list of at most 2000 shot IDs from the test collection ranked according to their likeliness of containing the concept.
- Across the years, systems submitted results for minimum 10 and up to 346 visual concepts.

Samples of concepts evaluated

3 Airplane
4 Airplane_Flying
9 Basketball
13 Bicycling
15 Boat_Ship
16 Boy
17 Bridges
25 Chair
31 Computers
51 Female_Person
54 Girl
56 Government_Leader
57 Greeting
63 Highway
71 Instrumental_Musician
72 Kitchen
74 Landscape
75 Male_Person
77 Meeting
80 Motorcycle
84 Nighttime
85 Office
95 Press_Conference
99 Roadway_Junction
101 Scene_Text
105 Singing
107 Sitting_down
112 Stadium
116 Teenagers
120 Throwing
128 Walking_Running
155 Apartments
163 Baby
198 Civilian_Person
199 Clearing
254 Fields
267 Forest
274 George_Bush
276 Glasses
297 Hill
321 Lakes
338 Man_Wearing_A_Suit
342 Military_Airplane
359 Oceans
434 Skier
440 Soldiers

Some automatic tagging results

Keyframes from the top 20 clips returned by a system looking for "boat_ship"

Content-based copy detection (2008 – 2011)

Example use case: *automatically determine whether a given video contains a (transformed) segment of any reference video (e.g., proprietary, illicit, etc.)*

Test Data:

- Create 201 base clips (2/3 containing reference video)
- Apply each combination of 8 video & 7 audio transformations to the base clips to create 11256 test clips (7*8*201)

System task:

- Given a 400 hour reference collection of Internet videos and 11256 test clips
- Determine for each test clip **whether** it contains reference video and if so, **where** that reference video begins and ends in the test clip.

Applications:

- Copyright control (e.g. MovieLabs)
- Business intelligence (advertisement tracking)
- Law enforcement investigations involving specific video

Transformations

Video

- Simulated camcording
- Picture in picture
- Insertions of pattern
- Strong re-encoding
- Change of gamma
- Decrease in quality
- Post production
- Combination of 3 randomly selected

Audio

- Nothing
- mp3 compression
- mp3 compression and multiband companding
- Bandwidth limit and single-band companding
- Mix with speech
- Mix with speech, then multiband compress
- Bandpass filter, mix with speech, compress

Surveillance Event Detection (2008 – 2017)

Two system tasks:

- **Retrospective SED (rSED)**: Given a textual description of an observable event of interest, automatically detect all occurrences of the event in a non-segmented corpus of video
- **Interactive SED (iSED)**: Given a textual description of an observable event of interest, at test time allow a searcher 25 minutes to filter incorrect event detections from the rSED task

Motivation:

- Streaming detection application on large, real-world data
- Naturally occurring events
- Human-centric event definition

CPNI
Centre for the Protection of National Infrastructure

Home Office
Scientific Development Branch

Homeland Security

NIST National Institute of Standards and Technology

Evaluation Source Data

iLIDS Imagery Library for Intelligent Detection Systems

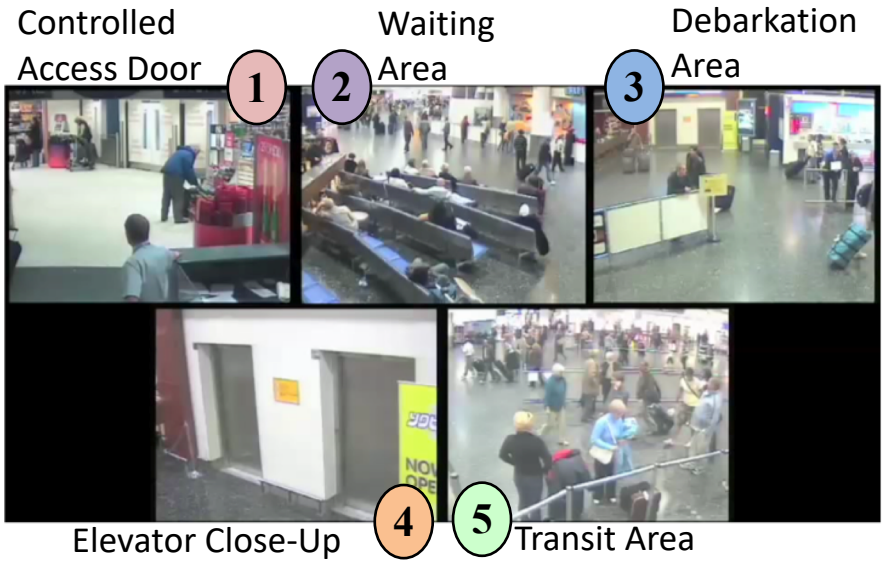
UK Home Office collected CCTV video from 5 camera views at a busy airport

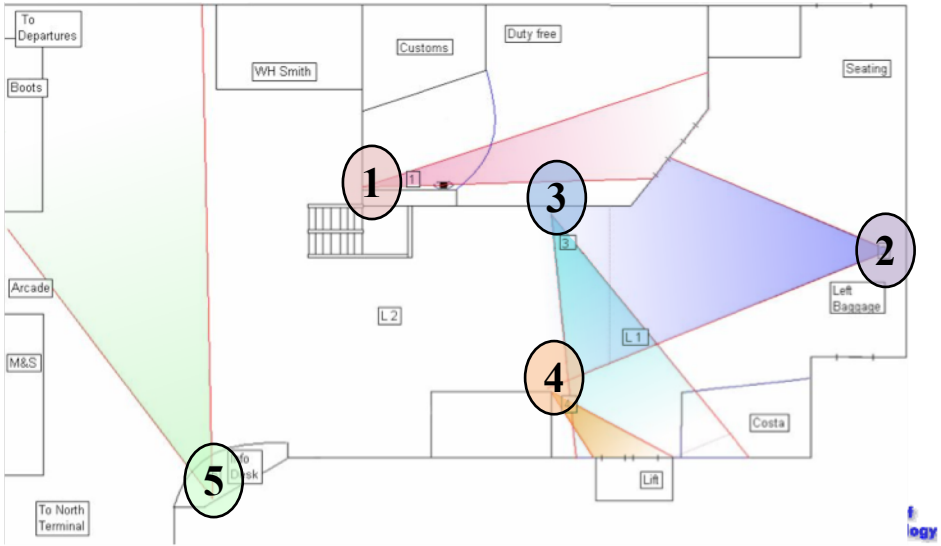
Development Set

- 100 hours of video
- 10 events annotated on 100% of the data

Evaluation Set

- "iLIDS Multiple Camera Tracking Scenario Training set"
- Subset of the ~50-hour iLIDS data set





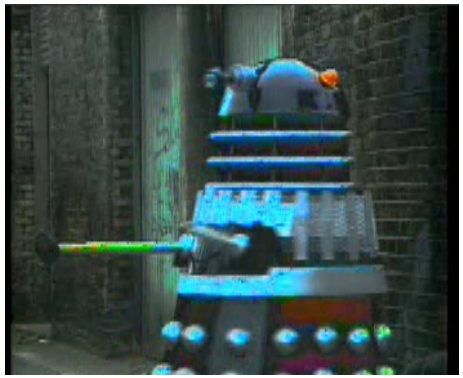
SED Events

Single Person events	
PersonRuns	Someone runs ← *Lowest frequency*
Pointing	Someone points ← *Highest frequency*
Single Person + Object events	
CellToEar	Someone puts a cell phone to his/her head or ear
ObjectPut	Someone drops or puts down an object
Multiple People events	
Embrace	Someone puts one or both arms at least part way around another person
PeopleMeet	One or more people walk up to one or more other people, stop, and some communication occurs
PeopleSplitUp	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame

iLIDS® Imagery Library for Intelligent Detection Systems

Instance Search Task (2010 – Present)

Example use case: *browsing a video archive, you find a video of a person, place, or thing of interest to you, known or unknown, and want to find more video containing the same target, but not necessarily in the same context.*



System task:

- Given a topic with:
  - 4 example images of the target
  - 4 ROI-masked images
  - 4 shots from which example the images came
  - a target type (OBJECT/LOGO, PERSON)
  - <topic title>
- Return a list of up to 1000 shots ranked by likelihood that they contain the topic target (person, object, location, Person X *in* location Y)
- **Automatic** or **interactive** runs are accepted

Data …

The BBC and the AXES project made **464 hours** of the BBC soap opera EastEnders available for research

- 244 weekly "omnibus" files (MPEG-4) from 5 years of broadcasts
- 471527 shots
- Average shot length: 3.5 seconds
- Transcripts from BBC
- Per-file metadata

Represents a "small world" with a slowly changing set of:

- People (several dozen)
- Locales: homes, workplaces, pubs, cafes, open-air market, clubs
- Objects: clothes, cars, household goods, personal possessions, pets, etc
- Views: various camera positions, times of year, times of day,

Use of fan community metadata allowed, if documented

Topic examples

Topic#: True positives:

99 494

A checkerboard band ...

106 243

Underground logo

101 1568

A Primus ... machine

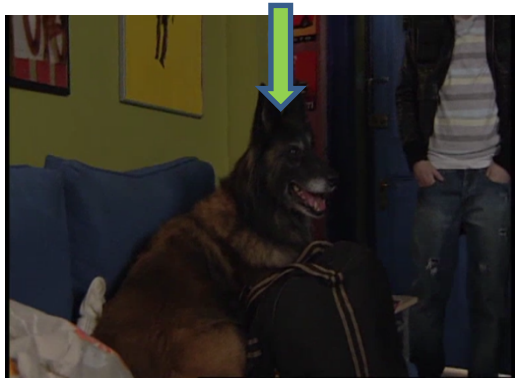
102 398

This large vase ...

115 277

This man

105 97



This dog, Wellard

Multimedia Event Detection (2010 – 2017)

Task:

- Quickly **find** instances of ***complex events*** in a large collection of ***Internet videos*** and **recount** the evidence in a search video for the event.

Motivation:

- There is a need for more capabilities to search videos beyond just human-generated metadata.
- Current prototype video search systems **weak on events with complex temporal structure** – not recognizable in a single image.
- Video analysts dealing with enormous volumes of video and changing search interests need "*index once – search repeatedly*" , content-based solution.

Example events :

- Making a cake, Assembling a shelter, Repairing an appliance, Grooming an animal, etc

Multimedia Event Detection (MED) Task

Flash Mob Gathering Event Kit

Definition: (text)

A coordinated large group of people assemble suddenly in a public place, perform a predetermined act to a surprised public, then disperse quickly

Explication: (text)

A flash mob is a group of people in a public place surprising the public by doing something unusual in a coordinated fashion. Flash mobs usually consist of people either suddenly starting to perform a ...

Evidential Description: (text)

- scene: indoor or outdoor, public place
- objects/people: a very large group of people, typically no objects involved
- activities: a wide range of activities can be performed, including dancing or singing in unison, ….
- audio: background music; sound that designates start/end of the flash mob activity; leader speaking to group of assembled flash mobbers

Illustrative Examples: (video)

- Positive instances of the event
- Clips "Related" to the event

Flash mob gathering positive example

HAVIC Data Resources

		Video clips	Video duration	2014 Change
Development Data	RESEARCH	10,000	314 hours	-
	10 Event Kits	1,400	74 hours	-
	Transcription	1,500	45 hours	-
Event Training Data	Event Background	5,000	146 hours	-
	40 Event Kits	6,000	270 hours	+53 hours
Test Data	MEDTest	27,000	849 hours	+11 hours
	KindredTest	14,500	687 hours	+11 hours
Evaluation Data	MED14Eval-Full	198,000	7,580 hours	+ 3,858 hours
	MED14Eval-Sub	33,000	1,244 hours	- 7 hours
	Total	244,000	9,911 hours	+3,922 hours

Strassel, S., Morris, A., Fiscus, J.G., Caruso, C., Lee, H., Over, P., Fiumara, J., Shaw, B., Antonishek, B. and Michel, M., 2012, May. Creating HAVIC: Heterogeneous Audio Visual Internet Collection. In *LREC* (Vol. 4, p. 2)

Concept Localization (2013 – 2016)

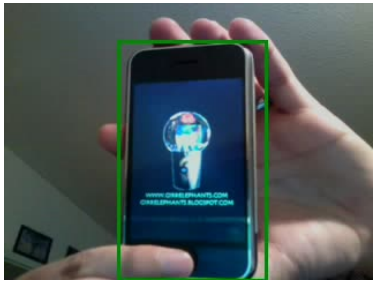
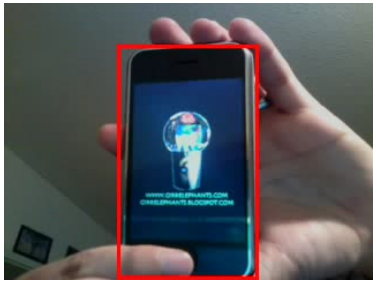
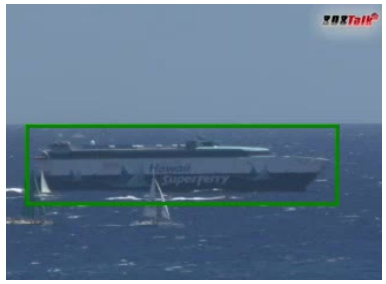
- **Task:**
  - Return the x,y coordinates of the upper left and lower right vertices of a bounding rectangle (in all I-frames of video shots) which contains all of the target concept and as little more as possible.
- Evaluated Spatial and Temporal localization
- **Target concepts:**
  - Airplane Anchorperson Boat_Ship Bridges
  - Bus Computers Motorcycle Telephones
  - Flags Quadruped Bicycling Boy
  - Baby Skier Running Explosion_fire
  - Dancing Instrumental_Musician Sitting_Down
  - Chair Hand

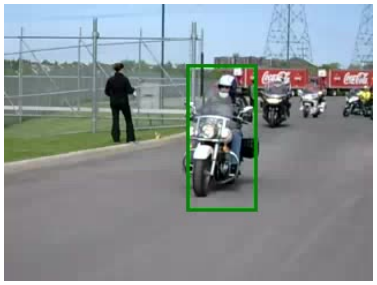


GT
Sys

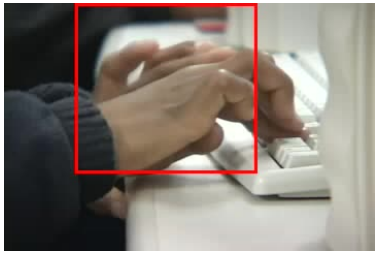
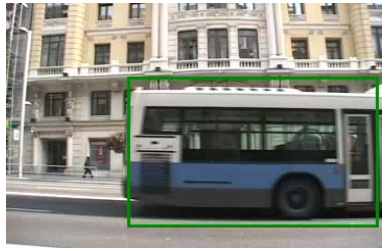




Samples of good localization





Some TRECVID impacts

Better ideas
Better systems
Economic benefits
Tech Transfer

More, better ideas

- Thornley, C. V., Johnson, A. C., Smeaton, A. F., & Lee, H. (2011). **The scholarly impact of TRECVid (2003–2009)**. Journal of the American Society for Information Science and Technology, 62(4), 613-627.
  - 310 workshop papers
  - 2073 peer-reviewed journal/conference papers (derived pubs)
- **2010 – 2017**
  - 346 workshop papers
  - 722 peer-reviewed ACM/IEEE pubs ONLY

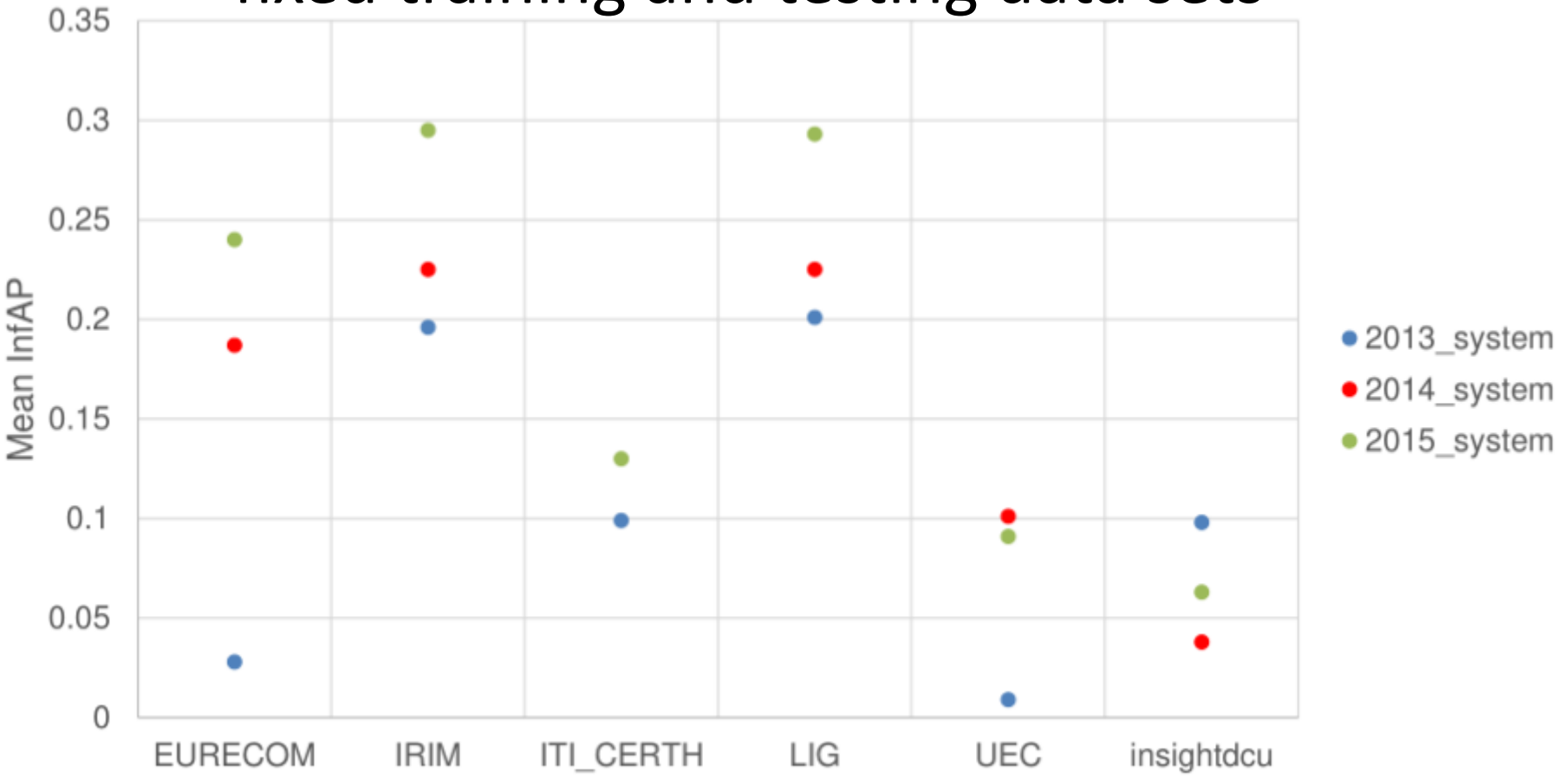
More, better systems

- Continuing improvement in feature detection (automatic tagging) in the University of Amsterdam's MediaMill system
  - Performance on 36 features doubled: 2006 –> 2009
    - Within domain (train and test) MAP 0.22 -> 0.41
    - Cross domains MAP 0.13 -> 0.27
  - In 2015, due to deep learning approaches, performance also similarly improved.

- You can only improve your performance if you consistently measure it!

("If you can not measure it, you can not improve it." – Lord Kelvin)

Measuring progress of Concept detection from 2013 to 2015 using fixed training and testing data sets



Economic impact: 2010 RTI International economic impact study of TREC/TRECVID

- RTI International (hired by NIST) undertook retrospective economic impact assessment of NIST's TREC and TRECVID program activities 1991 – 2009
- Stakeholders: IR researchers + Search system users
- Method:
  - Economic costs/benefits estimated under counterfactual scenario that TREC/TRECVID had not existed
  - Semi-structured interviews and a web survey
  - 404 respondents (93 based in US)
    - 30% at US-owned software or IR services companies (58% of total 2008 R&D expenditures on IR)
    - 66% at US universities (47% of total 2008 US university research expenditures)

"…for every $1 that NIST and its partners invested in TREC[/TRECVID], at least $3.35 to $5.07 in benefits accrued to IR [Information Retrieval] researchers" (http://trec.nist.gov/pubs/2010.economic.impact.pdf)

Tech transfer example

Euvision Technologies[*1] makes Amsterdam U. semantic indexing software commercially available as Impala *

Thu, 2 Feb 2012

Mr. Over:

We expect to sign our first paid licensing agreement next week. Licensee will be a system integrator who then makes the software available to all police departments in the Netherlands. Concepts to detect are nudity, babies, and children. Application is detection of child abuse in images/videos on confiscated computers/DVDs/tapes.

Your work will have impact on society, in a good way.

Kind regards,

Harro Stokman.

CEO Euvision Technologies,

M: +31 6 41 51 95 67

www.euvt.eu

Matrix II / Science Park 400

1098 XH Amsterdam

NetherlandsEuvision Technologies -/- Premier Visual Concept Detection

** Identification is not intended to imply recommendation or endorsement by NIST

[1] Acquired by QUALCOMM in 2014

Euvision technologies – customer applications *

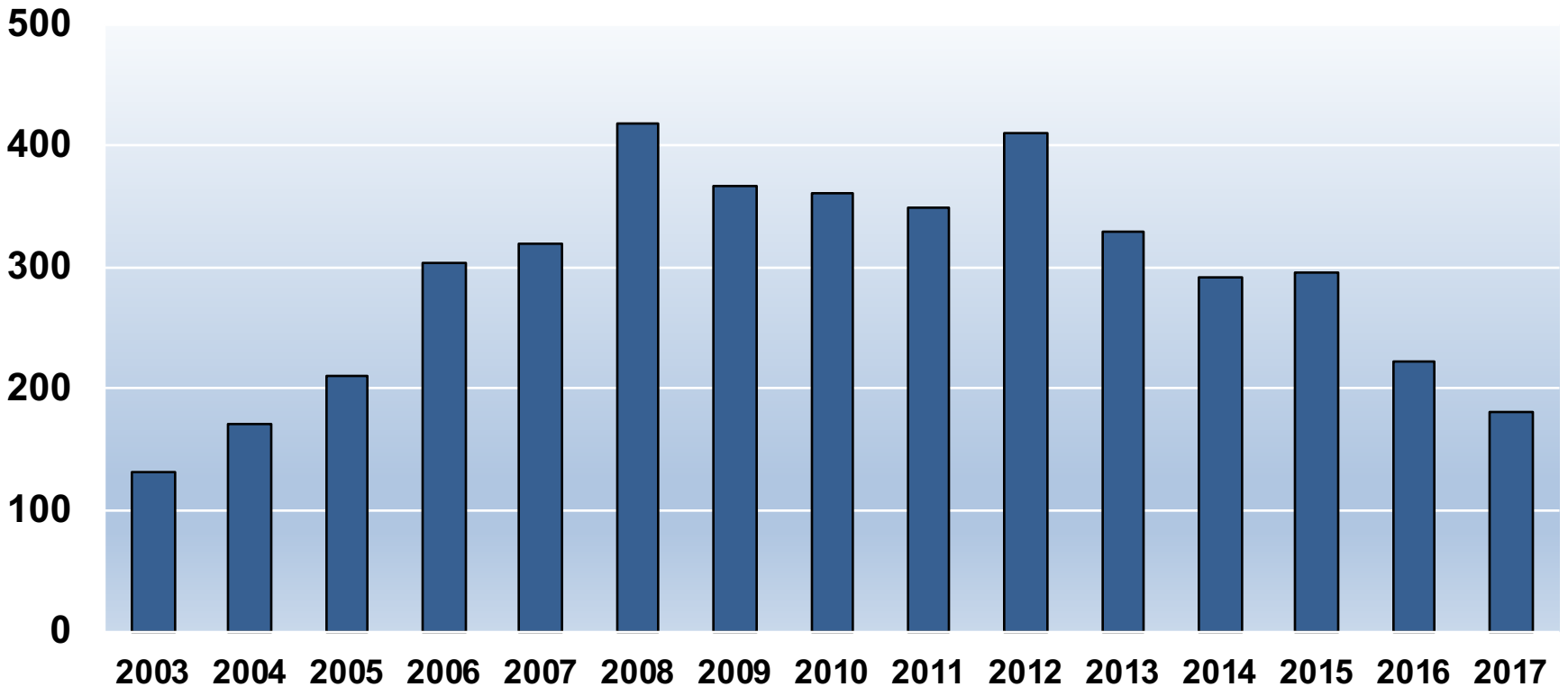
Gerrit Baarda, CEO of Ziuz (www.ziuz.com), says: “We have licensed Impala and integrated it into VizX2. VizX 2 is a total solution for analyzing video and photo material confiscated in investigations into sexual child abuse. Our clients love the new filtering technology. They find the illegal stuff faster, with decreased mental stress for the team.?

Toon Akkermans, CEO of NCIM (www.ncim.com), says: "We have integrated Impala in our Forensic Dashboard. This Dashboard sits on data of the Dutch Forensic Institute (NFI/Xiraf). In several E-discovery cases, we tried to find documents containing invoices in a big pile of data. Existing text based search found a few: only the ones that were tagged as an invoice. Impala found the rest, hidden in huge set of images. Today, we therefore start with Impala based search."

* Identification is not intended to imply recommendation or endorsement by NIST

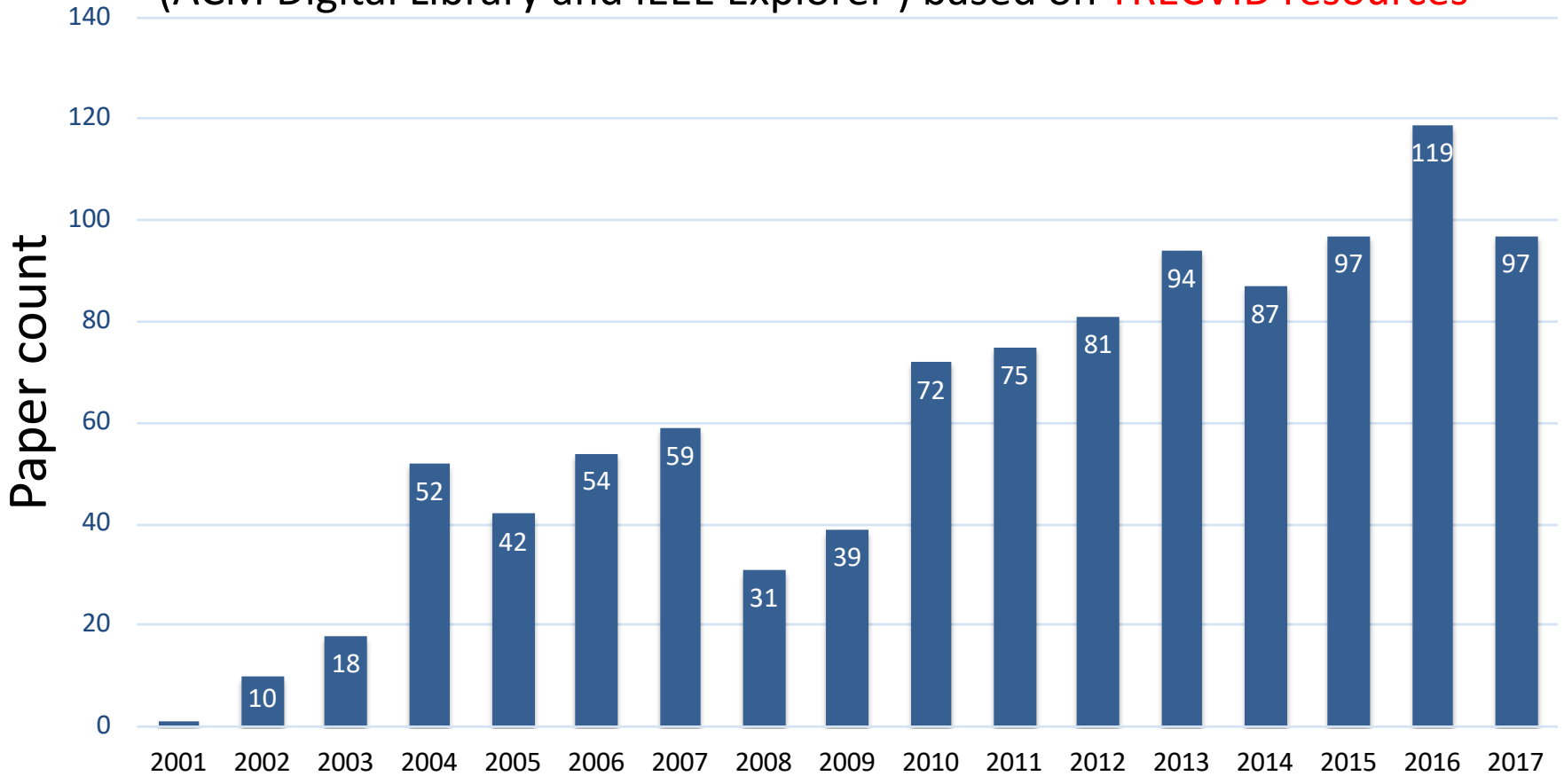
Where are we now?

Workshop paper author count by year



TRECVID Bibliography

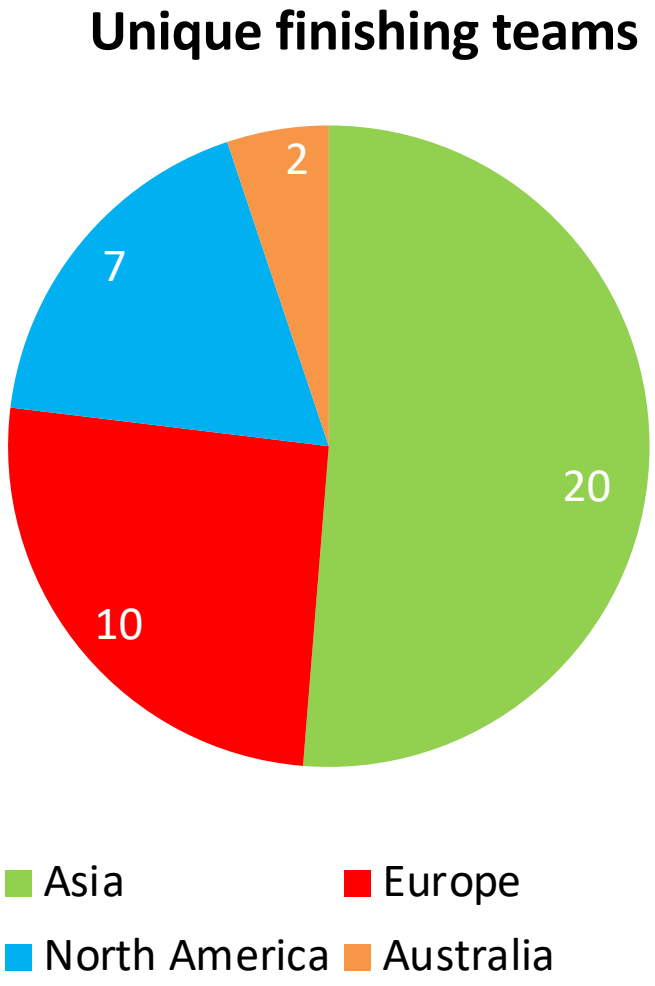
Partial bibliography of peer-reviewed journal and conference papers (ACM Digital Library and IEEE Explorer) based on TRECVID resources



http://www-nlpir.nist.gov/projects/trecvid/trecvid.bibliography.txt

TV2017 Finishers* (35 out of 76)

Teams Finished	Task code	Task name
7	SED	Surveillance event detection
10	AVS	Ad-hoc Video Search
8	INS	Instance search
6	MED	Multimedia event detection
3	LNK	Video hyperlinking
16	VTT	Pilot task (Video_to_Text)



*A team that submitted at least 1 run

Observations, questions …

- **One solution will not fit all.** Investigations/discussion of video search must be related to the searcher‘s specific needs/capabilities/history and to the kinds data being searched.
- **The enormous and growing amounts of video require extremely large-scale approaches** to video exploitation. Much of it has little or no metadata describing the content in any detail.
  - **400 hrs** of video are being uploaded on YouTube per **minute** (as of 11/2017)
  - **1 billion hrs** of video content are watched **per day** (as of 2/2017)
- TREVCID participants have explored some automatic approaches to tagging and use of those tags in automatic and interactive search systems. **Much has been learned, some results may already be useful, but still a lot of work need to be explored.**

Observations, questions …

Within the focus of TRECVID experiments …

- Multiple information sources (text, audio, video), each errorful, can yield better results when **combined** than used alone…
- A human in the loop in search still makes an enormous difference.
- Text from speech via automatic speech recognition (ASR) is a powerful source of information but:
  - Its usefulness varies by video genre
  - Not everything/one in a video is talked about, “in the news"
  - Audible mentions are often offset in time from visibility
  - Not all languages have good ASR
- Machine learning approaches to tagging
  - yield seemingly useful results against large amounts of data when training data is sufficient and similar to the test data (within domain)
  - but will they work well enough to be useful on highly heterogeneous video?

Observations, questions …

Within the focus of TRECVID experiments …

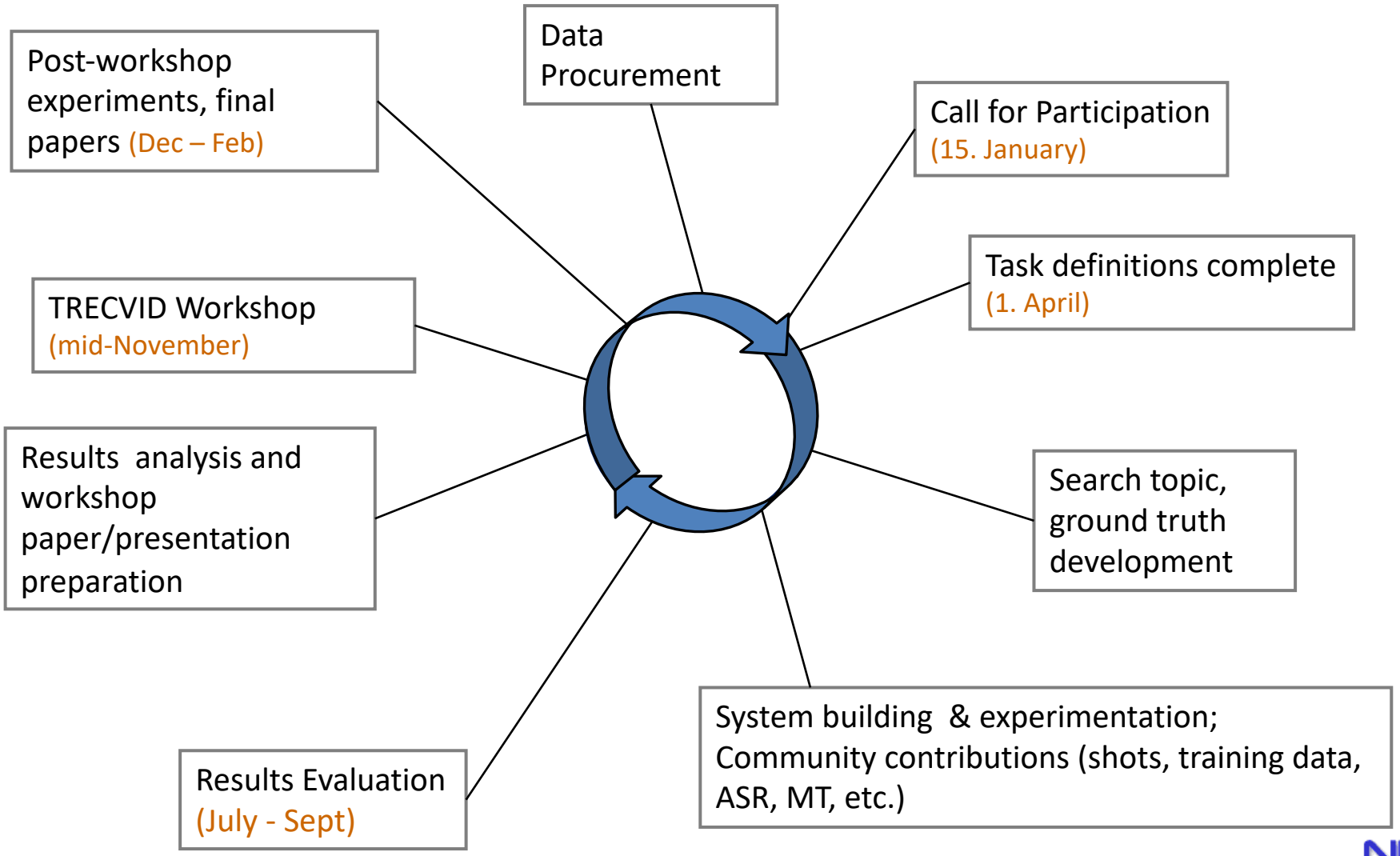
- Searchers (experts and non-experts) will use more than text queries if available: concepts, visual similarity, temporal browsing, positive and negative relevance feedback,… www.videobrowsershowdown.org/

Observations, questions …

Within the focus of TRECVID experiments …

- Processing video using a sample of more than one frame per shot, yields better results but quickly pushes common hardware configurations to their limits
- TRECVID systems have been looking at combining automatically derived and manual-provided evidence in search :
  - Internet Archive video will provide titles, keywords, descriptions
  - Where in the Panofsky hierarchy are the donors' descriptions? If very personal, does that mean less useful for other people?
- Need observational studies of **real** searching of various sorts using current functionality and identifying unmet needs

TRECVID Yearly Cycle



New in 2019

- New web video dataset
  - Vimeo creative commons creations (V3C1)
    - ~1000 hrs
    - 1.3TB Size
    - 7475 videos, 1M video shots
    - To support the Ad-hoc video search track for multiple years.
- New instance search query types (BBC Eastenders)
  - Find person X doing action Y
    - e.g. Find Stacey {Eating, drinking, sleeping, hitting someone}

ActEV (Activities in Extended Video)

- ActEV Leaderboard Evaluation
- ActEV Prize Challenge
- WACV'19 workshop "Human Activity Detection in Multi-Camera Video Streams"
- Please visit https://actev.nist.gov/

WACV'19 workshop: Human Activity Detection in Multi-Camera Video Streams

- workshop at the IEEE Winter Conf. on Applications of Computer Vision (WACV), Hawaii, January 7-11, 2019
- three invited talks from experts in the field
- four talks from the best performers at the ActEV evaluation
- the rest of the performers will be invited to present their work as a poster
- four regular papers will be selected after review

For more information

- trecvid.nist.gov
- Resources available
  - Annual tasks guidelines
  - Video datasets
  - Ground truth/judgments data
  - Topics/queries
  - Evaluation/scoring software
  - Publications of participating teams
  - Archived results and run submissions of past teams

Hope to see you at TRECVID 2019!

Thank You ☺