Open Challenges of Interactive Video Search and Evaluation

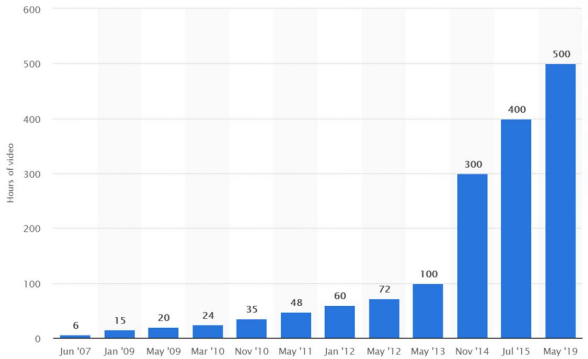
Jakub Lokoc, Klaus Schoeffmann, Werner Bailer,
Luca Rossetto, and Björn Þór Jónsson

ACM Multimedia 2022, Tutorial, Oct 10th, 2022

Introduction

Why Content-Based Search in Video?

- **Video as ubiquitous media**
  - to store/memorize, document, demonstrate, explain, …
- **Videos always and everywhere!**
  - Entertainment and commercials
  - Distance learning and documentation of events
  - Social gaming (screencasts)
  - Personal videos (hobbies, vacation, family, kids, …)
  - Sports documentation and analysis (e.g., GoPro)
  - Product usage instructions (e.g., furniture)
  - Surveillance (buildings, places, street, …)
  - Health care and medicine (endoscopic procedures)
  - Lifelogging



hours of video uploaded to YouTube every minute (May 2019), source: statista 2021

- **However, while recording is easy, search is not!**
  - most videos are not annotated
  - hence, automatic search for relevant scenes/clips is only possible, if the content is automatically analyzed and indexed

No Problem, We Have Neural Nets!

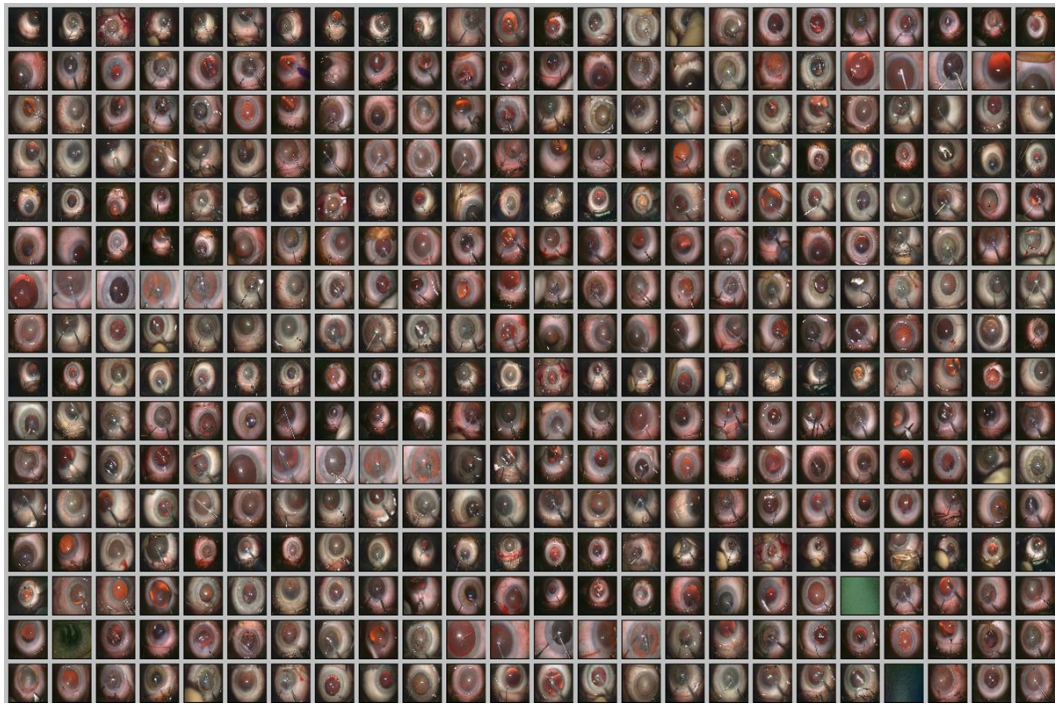
- **Simply apply a CNN and object detector (e.g., YOLO) to every frame!?**
  - might work for some videos and domains
  - for the majority of cases, however, this does not work at all!
    - **too low performance in general (still)**
    - **too specific classes for the domain**
    - **too many classes for common users**
    - **too slow search (without proper indexes)**
    - **too many results (without appropriate visualization)**

"Strike with a Pose"

No Problem, We Have Neural Nets!

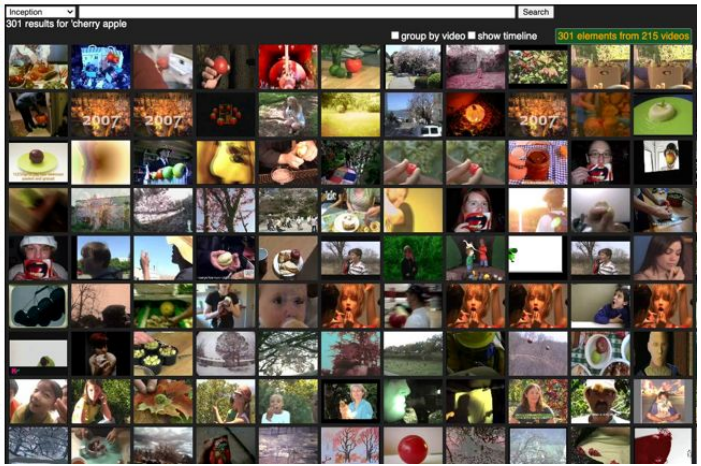
How would you search for this video?

in a monthly archive of 480
cataract surgery videos?



No Problem, We Have Neural Nets!

- **We have detected all 21,841 ImageNet classes in all frames! - YESSS👍**
- **BUT**
  - which concept should I choose?
  - what exactly is the difference between …
  - there is green pepper, spinach, lettuce, carrot, red cabbage, savoy cabbage, head cabbage, turnip cabbage, broccoli, etc. – this is so specific, why I can't search just ***green vegetables***?



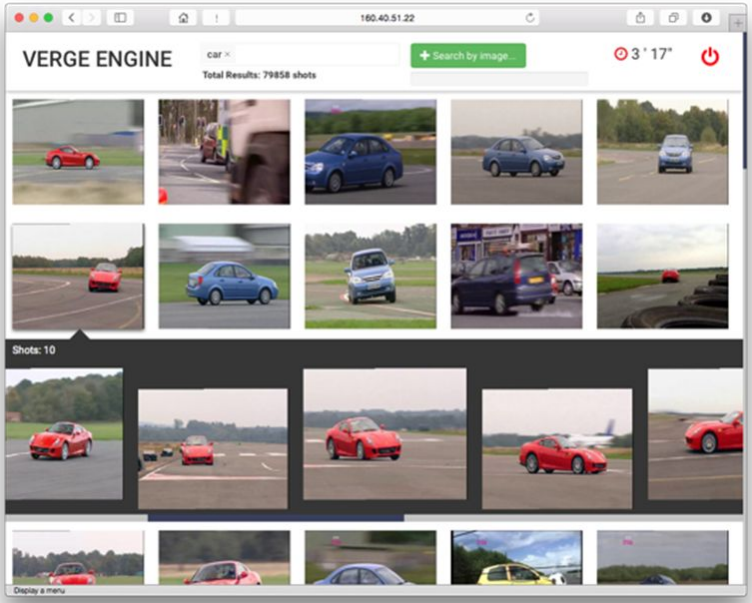
Automatic vs. Interactive Search

- **Beyond words and language**
  - how does a subway musician look like?
  - I cannot find the correct term in English…
  - which is the best concept for images like these?

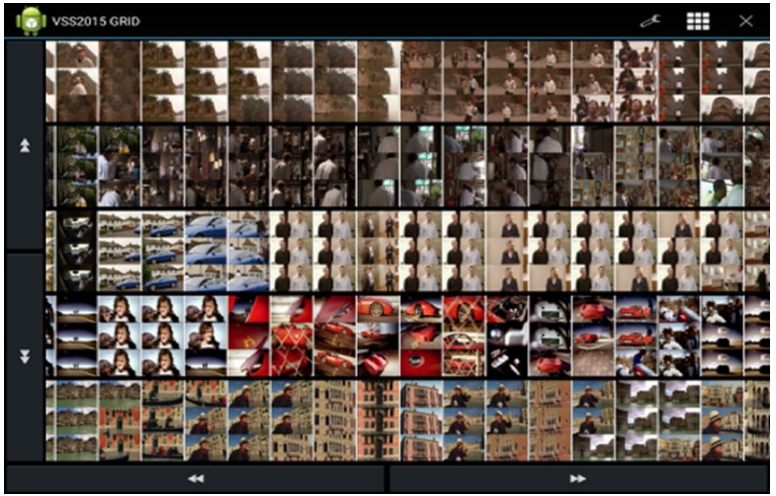
"An image tells a thousand words."

Automatic vs. Interactive Search

Example from the Video Browser Showdown 2015: (search in 100h video)



System X: shot and scene detection, concept detection (SIFT, VLAD, CNNs), similarity search.



System Y: tiny thumbnails only, tablet interface, powerful user.
Outperformed system X and was finally ranked 3rd!

Typical Types of Search

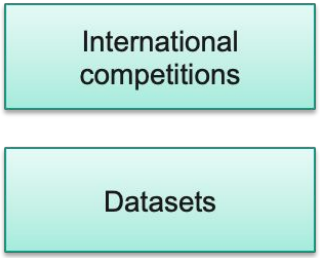
- **search-by-text**
  - enter keywords to match with available/extracted text (e.g., metadata, OCR, ASR, concepts, objects, …)
- **search-by-concept**
  - show results for a specific class/category (e.g., from ImageNet, MS COCO, etc.)
- **search-by-example**
  - provide example image/clip/sound
- **search-by-filtering**
  - filter content by some metadata or content feature (length, recording time, color, motion, genre, etc.)
- **search-by-sketch**
  - provide sketch of image or scene (e.g., color signatures, object sketch, motion sketch etc.)
- **search-by-relevance feedback**
  - start with random content, suggest content
  - provide iterative relevance feedback and repeat
- **search-by-browsing**
  - start by looking around, browsing content, and narrowing down search area
  - needs supportive visualization (e.g., similarity arrangement, video summaries, etc.)
- **search-by-exploration**
  - open and skim/watch specific videos
  - browse, navigate, filter, etc.
  - combine other features (see list above)

Search in multimedia content (particularly video) is a highly interactive process!

Users want to look around, try different query features, inspect results, refine queries, and start all over again!

Fair and Reproducible Evaluation

- **Interfaces are inherently developed for humans!**
- **Every user might interact differently**
  - different culture, knowledge, preferences, experiences, ..
  - even the same user at a different time
- **Video search interfaces need to be evaluated with real users…**
  - no simulations!
  - find out how well users perform with a real system!
- **…and with real data!**
  - real videos “in the wild” (e.g., V3C dataset)
  - actual queries that make sense in practice
- **Comparable and reproducible evaluations!**
  - same data, same query, same device/display, same conditions!
  - user studies and challenges/campaigns (e.g., TRECVID, VBS, LSC)



Video Browser Showdown (VBS)

- **Live evaluation platform for interactive video search**
  - content-based video retrieval at large scale
  - evaluates several types of search (KIS, AVS)
- **Competitive setup for comparable and fair evaluation**
  - same queries, same dataset, same conditions, at the same time!
  - direct competition with other systems (for several hours)
  - reveals both search performance and usability
    - **sophisticated scoring**
    - **expert and novice session**
    - **general domain and highly redundant domain *NEW***
- **Entertaining annual event**
  - part of the Welcome Reception at the MMM conference
  - showcases state-of-the-art video retrieval

Video Browser Showdown (VBS)

Evaluation Server (Projector Wall)

Judge

Judge

Moderator

Team 1
Team 2
Team 3
Team 4
Team 5
Team 6
Team 7
Team 8
Team 9

Audience

Audience

Audience

Research Goals of the VBS

- **Provide a platform for *comparable evaluation* of video search tools**
  - As alternative to user studies and user simulations
  - Same queries, same dataset, same conditions, at the same time!
  - Standardized interaction logging
- **Push research on video content search tools that are**
  - **Highly interactive**
  - **Efficient in terms of search time and accuracy**
  - **Flexible in terms of queries**
  - **Easy to use**

	KIS		AVS
	visual	textual	
Experts	x	x	x
Novices	x		x

- Luca Rossetto, Ralph Gasser, Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, Tomas Soucek, Phuong Anh Nguyen, Paolo Bolettieri, Andreas Leibetseder, and Stefanos Vrochidis. 2020. Interactive Video Retrieval in the Age of Deep Learning – Detailed Evaluation of VBS 2019. IEEE Transactions on Multimedia (TMM)

VBS is a Collaborative Effort!

VBS Organization Team

Klaus Schoeffmann
Klagenfurt University

Werner Bailer
Joanneum Research

Jakub Lokoc
Charles University

Cathal Gurrin
Dublin City University

Active Contributors and Collaborators

Luca Rossetto
University of Zurich

Ralph Gasser
University of Basel

Loris Sauter
University of Basel

George Awad
NIST/USA

- 30 unique teams from 21 countries (nearly 200 unique authors)!
- 87 different systems so far
- dozens of people who contributed as judges (and to server software)

Video Browser Showdown (VBS)

Team	Organization	Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
ITEC/diveXplore	**Institute of Information Technology, Klagenfurt University**	**Austria**	x	x	x	x	x	x	xx	x	x	xxx	x
IMOTION/vitrivr(-VR)	**University of Basel**	**Switzerland**				x	xx	x	x	x	x	xx	xx
VideoGraph	University of Zurich	Switzerland										x	x
VIRET/SOM-Hunter	**Charles University, Prague**	**Czech Republic**			x	x	x	x	x	x	xx	xxx	x
VIREO	Hong Kong City University	Hong Kong						x	x	x	x	x	x
VISIONE	Institute of Information Science and Technologies, Pisa	Italy								x		x	x
VERGE	Information Technologies Institute/CERTH, Thessaloniki	Greece			x	x	x	x	x	x	x	x	x
NII-UIT/VNUHCM	**(NII, Tokyo / University of Information Technology), VNUHCM, Ho Chi Min City**	**Japan / Vietnam**		x	x	x		x	x		x		x
ViRMA	IT-University of Copenhagen	Denmark											x
V-FIRST	Vietnam National University / Dublin City University	Vietnam / Ireland											x
HTW/vibro	**HTW Berlin, University of Applied Sciences**	**Germany**				x	x		x			x	x
Exquisitor	IT University of Copenhagen, University of Amsterdam	Denmark / Netherlands									x	x	x
IVIST	Image and Video Systems Lab, KAIST, Daejeon	South Korea									xx	x	x
DCU/VideoFall & AVSeeker	Dublin City University / Vietnam National University	Ireland / Vietnam	x	x	x	x	x					x	xx
-	University of Southern California, German Jordanian University	USA / Jordan										x	
NECTEC	NECTEC, Pathum Thani	Thailand							x				
JR	JOANNEUM RESEARCH, Graz	Austria	x	x	x		x						
UU	Utrecht University	Netherlands				x	x						
NUS	National University of Singapore	Singapore	x				x						
VideoCylce	University of Mons	Belgium		x									
FRTRD	France Telecom Research & Development, Beijing	China		x									
OVIDIUS	artemis Department, Télécom SudParis, Evry Cedex	France	x										
ISYS	Institute of Information Systems, Klagenfurt University	Austria	x										
UPC	Polytechnic University of Catalonia, Barcelona	Spain	x										

Video Browser Showdown (VBS)

Distributed Retrieval Evaluation Server (DRES) - Best Demo @ MMM2021

Works fully virtual as well!
(VBS 2021 + 2022)

Many thanks to: Luca Rossetto, Ralph Gasser, Loris Sauter

→ see later

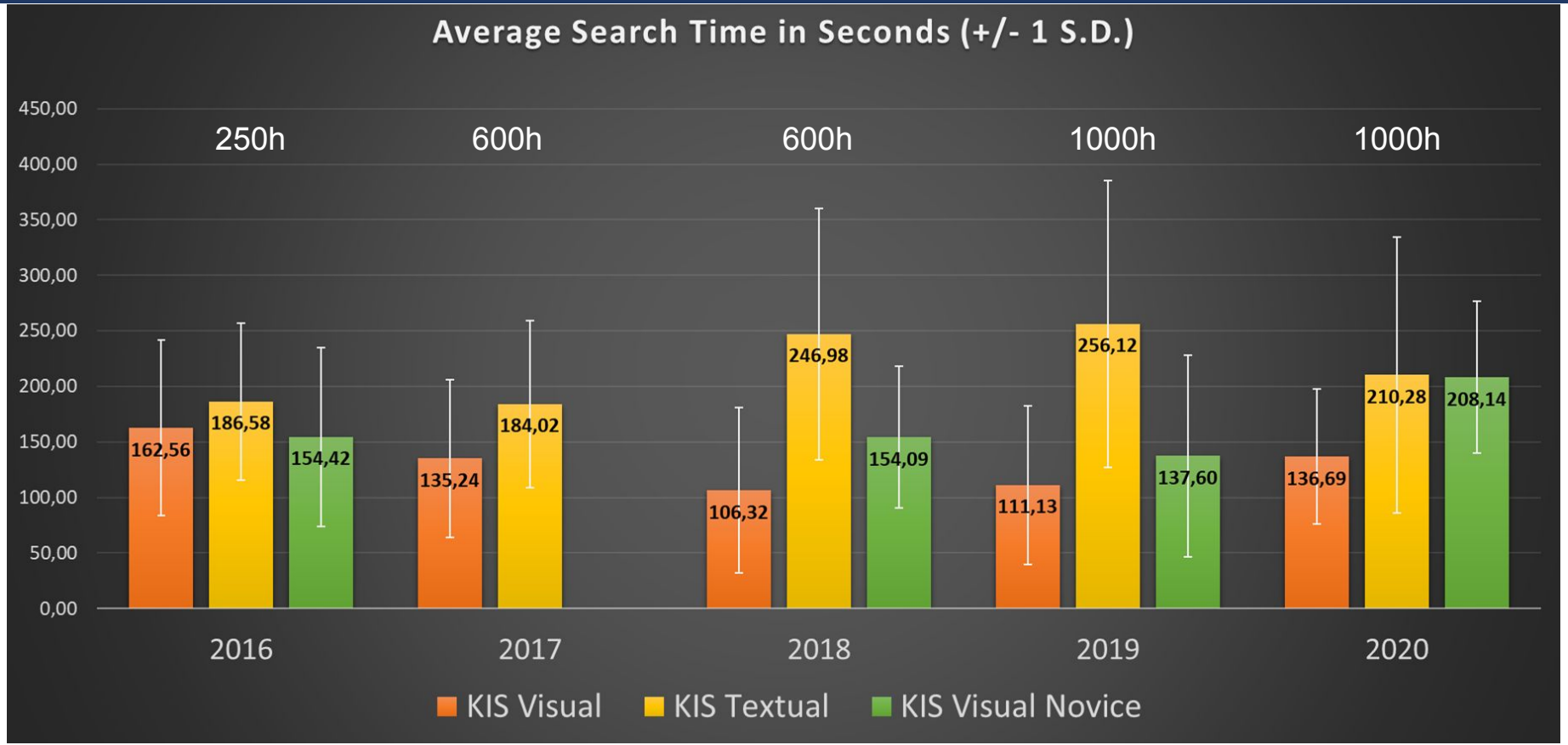
VBS Datasets

- **2012 (KIS in single videos)**
  - 30 video, 38 hours, visual (v)
- **2013 (KIS in single videos)**
  - 10 videos, 10 hours, visual (v) + textual (t)
- **2014 (KIS in single videos & collection)**
  - 76 videos, 30 hours for collection search, v + t
- **2015 (KIS in collection)**
  - 153 videos, 100 hours, v + t
- **2016 (KIS in collection)**
  - 442 videos, 250 hours, v + t
- **2017-2018 (KIS and AVS in collection)**
  - 4573 videos, 600 hours, v + t
  - AVS partly from TRECVID
- **2019-2021 (KIS and AVS in collection)**
  - 7475 videos, 1000 hours, v + t
  - 1.08 million segments
- **2022 (KIS and AVS in collection)**
  - 17235 videos, 2300 hours, v + t

Data source:

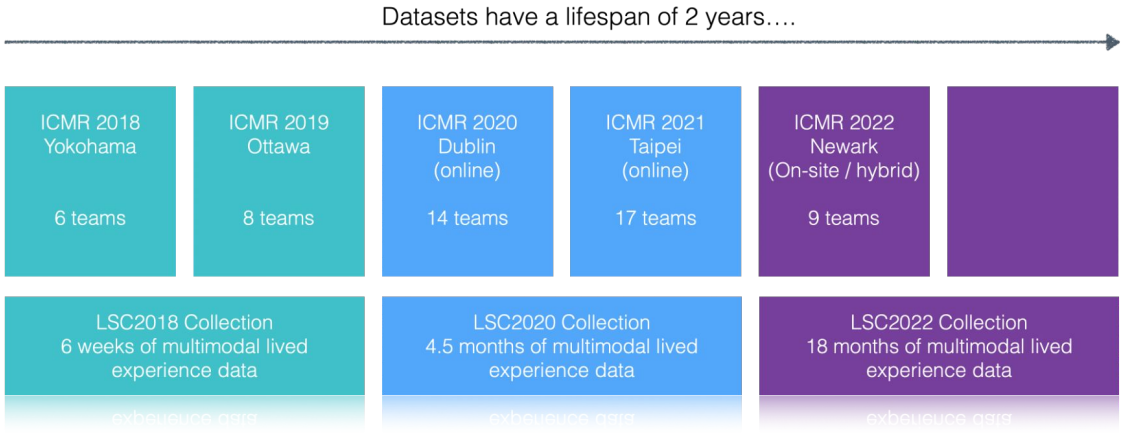
- 2012-2014: EBU SCAIE (and NHK), TOSCA-MP EU project
- 2015-2016: BBC MediaEval (Search & Hyperlinking task)
- 2017-2018: IACC.3 (Internet Archive Creative Commons)
- 2019-2021: V3C1 (Vimeo Creative Commons Collection)
- 2022-: V3C1 + V3C2 with about 2300 hours of content

Video Browser Showdown



Lifelog Search Challenge

- **Similar to the VBS but for multimodal lifelog data**
  - started at ICMR 2018 in Yokohama
  - has run annually since then
- **Time as a key factor**
  - something happens before/after something else
  - many visually similar items might exists, but context makes a difference
- **Smaller datasets to reduce the entry barrier**



Lifelog Search Challenge

0, 30, 60, 90, 120, 150... 300 seconds

I see Steve Wozniac…

I see Steve Wozniac on a wall of portraits.

I see Steve Wozniac on a wall of portraits. The wall was a brick wall with a door and large heater.

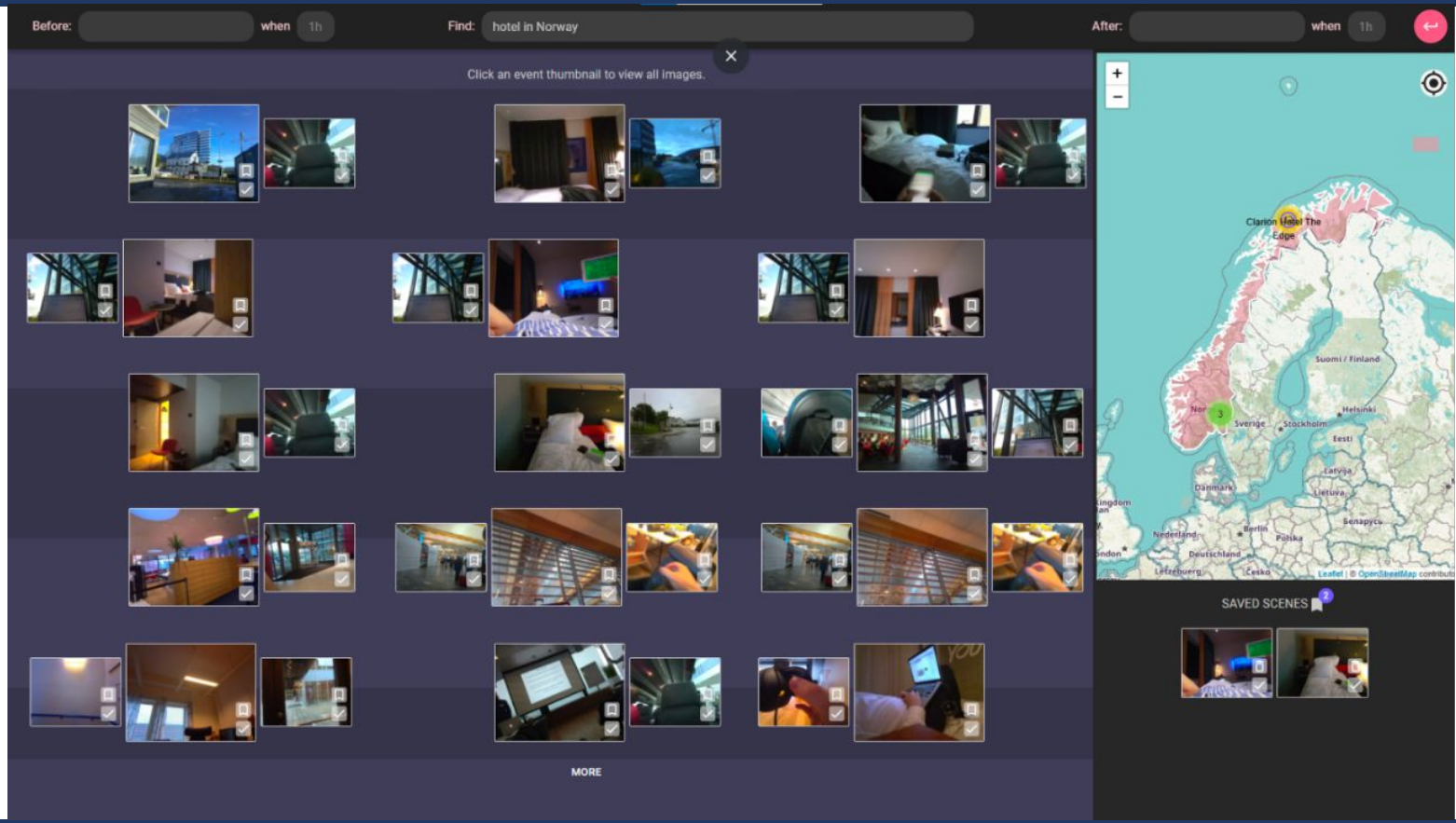
I see Steve Wozniac on a wall of portraits. The wall was a brick wall with a door and large heater. I was speaking to an audience before seeing the photos.

I see Steve Wozniac on a wall of portraits. The wall was a brick wall with a door and large heater. I was speaking to an audience before seeing the photos. I left by driving back to work.

I see Steve Wozniac on a wall of portraits. The wall was a brick wall with a door and large heater. I was speaking to an audience before seeing the photos. I left by driving back to work. It was in 2015 in March on a Wednesday.

Lifelog Search Challenge

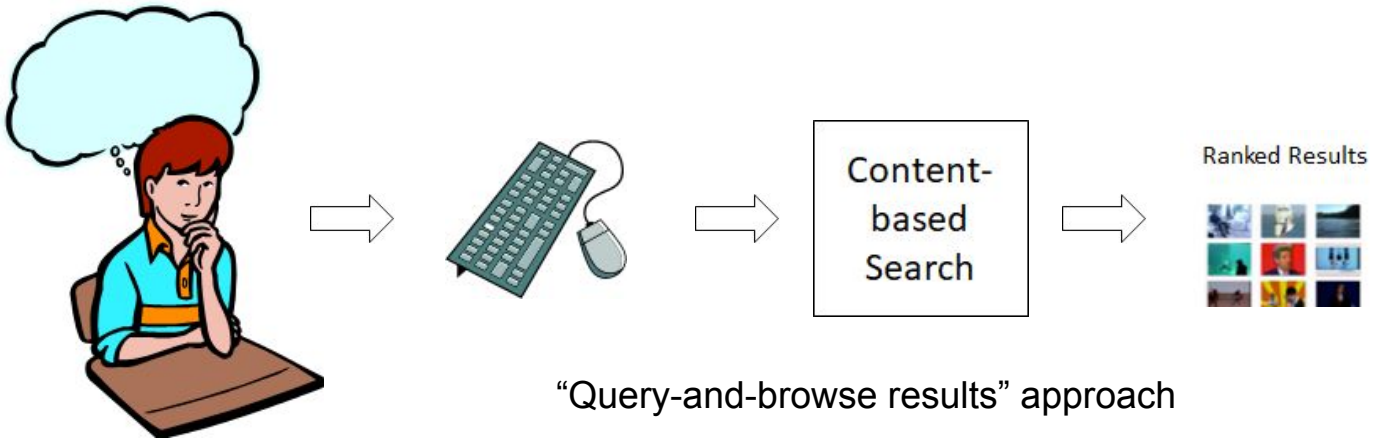
Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2020. Myscéal: An Experimental Interactive Lifelog Retrieval System for LSC'20. In Proceedings of the Third Annual Workshop on Lifelog Search Challenge (LSC '20). Association for Computing Machinery, New York, NY, USA, 23–28.



Tasks & Challenges

Fully automatic video retrieval

- **Works well if**
  - Users can properly describe their needs
  - System understands search intent of user
  - Content features can sufficiently describe visual content
  - Computer vision (i.e., CNNs) can accurately detect semantics
- **Unfortunately, in real-world rarely true**



“Query-and-browse results” approach

How interactive video retrieval differs

- **Assume a smart and interactive user**
  - That knows about the challenges and shortcomings of simple querying
  - But might also know how to circumvent them
  - Could be a digital native!
- **Give him/her full control over the search process**
  - Provide many query and interaction features
    - **Querying, browsing, navigation, filtering, inspecting/watchi**
- **Assume an iterative/exploratory search process**
  - **Search - Inspect - Think - Repeat**
  - "*Will know it when I see it*"
  - Could include many iterations!
  - Instead of "query-and-browse results"

Evaluation of Interactive Video Retrieval

- **Interfaces are inherently developed for human users**
- **Every user might be different**
  - Different culture, knowledge, preferences, experiences, ...
  - Even the same user at a different time
- **Video search interfaces need to be evaluated with real users...**
  - No simulations!
  - User studies and campaigns (TRECVID, MediaEval, VBS, LSC)!
  - Find out how well users perform with a specific system
- **...and with real data!**
  - Real videos “in the wild” (e.g., IACC.1 and V3C dataset)
  - Actual queries that would make sense in practice
  - Comparable evaluations (same data, same conditions, etc.)

Overview of Evaluation Approaches

- **Qualitative user study/survey**
  - Self report: ask users about their experience with the tool, thinking aloud tests, etc.
  - Using psychophysiological measurements (e.g., electrodermal activity - EDA)
- **Log-file analysis**
  - Analyze server and/or client-side interaction patterns
  - Measure time needed for certain actions, etc.
- **Question answering**
  - Ask questions about content (open, multiple choice) to assess which content users found
- **Indirect/task-based evaluation (Cranfield paradigm)**
  - Pose certain tasks, measure the effectiveness of solving the task
  - Quantitative user study with many users and trials
  - Open competition, as in VBS, LSC, and TRECVID

Properties of Evaluation Approaches

- **Availability and level of detail of ground truth**
  - None (e.g., questionnaires, logs)
  - Detailed and complete (e.g., retrieval tasks)
- **Effort during experiments**
  - Low (automatic check against ground truth)
  - Moderate (answers need to checked by human, e.g. live judges)
  - High (observation of or interview with participants)
- **Controlled conditions**
  - All users in same room with same setup (typical user-study) vs. participants via online survey
- **Statistical tests!**
  - We can only conclude that one interactive tool is better than the other, if there is statistically significant proof
    - **Tests like ANOVA, t-tests, Wilcoxon-signed rank tests, …**
    - **Consider prerequisites of specific test (e.g., normal distribution)**

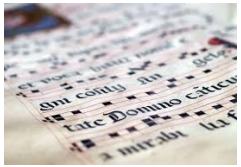
Practicality of Evaluation Approaches

- **Task-based approaches have a number of practical advantages**
  - ground truth can be reused
  - ground truth can be defined before (in theory, with limits …)
  - automatic assessment (if we have all the ground truth)
  - thus repeatable
  - objective results to chosen level of detail

Task Types: Introduction

- **Searching for content can be modelled as different types of tasks**
- **Task serves as a laboratory model of a real-world situation with an information need for multimedia data**
  - Isolate: consider one or few steps from a process
  - Standardise: create a framework of controlled conditions (applicability of evaluation metrics, repeatability)
  - Simplify: reduce complexity of the setting: to limit the number of variables
  - Task design choices impact dataset preparation, annotations (effort!), evaluation methods and the way to run the experiments

Task Types: Overview



Example *Visual cue* *Aural cue* *Music scores* *Textual* *None*

Specific item → **Query**

Result set

Class

KIS by example

KIS

QbE

Retrieval

Exploration

Target

Task category space dimensions

A_{CI}: **Number of correct items satisfying a search need in a dataset**

A_{SI}: **Requested number of submitted correct items**

A_{PM}: **Search need presentation modality**

A_{PT}: **Search need presentation timing**

A_{PQ}: **Search need presentation quality**

A_{DC}: **Data collection**

A_{TL}: **Time limit**

A_{US}: **User skills**

A_{NU}: **Number of operating users**

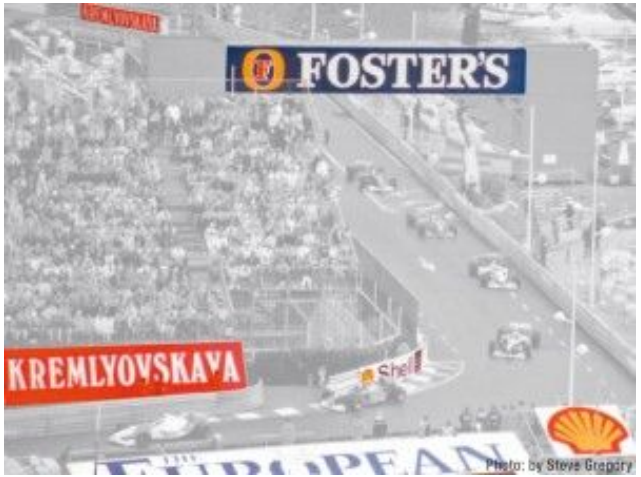
A_{QM}: **Quality measure**

Table 2. VBS'21 task categories represented as vectors of the task space. For each category, the value in an axis column presents the currently used axis option (specified in Section 3). Due to the virtual conference setting, only expert users were participating.

Task Name	A_{CI}	A_{SI}	A_{PM}	A_{PT}	A_{PQ}	A_{DC}	A_{TL}	A_{US}	A_{NU}	A_{QM}
Visual KIS	1	1	1	2	1	1	1	1	2	1, 3
Textual KIS	1	1	2	3	2	1	1	1	2	1, 3
Ad-hoc search	3	3	2	2	2	1	1	1	2	2, 3

Task Types: Query by Example

- User holds a digital representation of a relevant example of the needed information
- Example or its features can be sent to system
- User does not need to translate example into query representation
- e.g., trademark/logo detection



Task Types: Known Item Search (KIS)

- **User sees/hears/reads a representation of the needed information**
  - Used in VBS & LSC
- **Representation of exactly one relevant item/segment in content set**
- **Models cases where the user has a memory of content to be found**
- **User must translate the representation to query methods supported by the system**
  - The complexity of this translation depends significantly on the modality
  - e.g., visual is usually easier than textual, which leaves more room for interpretation
  - Relation of/to content is important
  - e.g. searching in own life log media vs. searching in media collection on the web

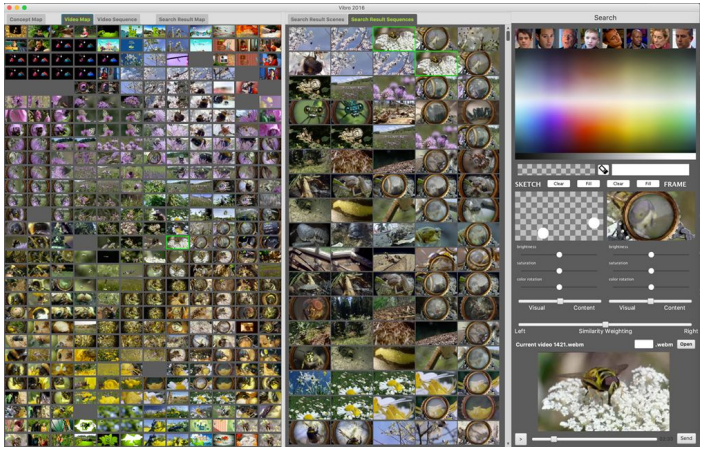
"on a busy street"

Task Types: Retrieval

- **User sees/hears/reads a representation of the needed information**
- **Representation of a broader set/class of relevant items/segments**
  - cf. TRECVID AVS task
- **Models cases where the user has a memory of the type of relevant content**
- **Similar issues of translating the representation like for KIS, but due to broader set of relevant items the correct interpretation of textual information is a less critical issue**
- **Raises issues of what is considered within/without scope of a result set**
  - e.g., partly visible, visible on a screen in the content, cartoon/drawing versions, …
  - TRECVID has developed guidelines for annotation of ground truth

Task Types: Exploration

- User does not start from a clear idea of the information need
- Browsing and exploring may lead to identifying useful content
- Reflects a number of practical situations, but very hard to evaluate
- No known examples of such tasks in benchmarking campaigns due to the difficulties with evaluation



Demo: https://www.picsbuffet.com/

Task Design is About Trade-offs: Aspects to consider

- **Tasks shall**
  - model real-world content search problems, in order to assess whether tools are usable for these problems
  - set controlled conditions, to enable reliable assessment
  - be repeatable, to compare results from different evaluation sessions
  - avoid bias towards certain features or query methods

many real world problems involve very fuzzy information needs	well defined queries are best suited for evaluation
users remember more about the scene when they start looking through examples	information in the task should be provided at defined points in time
during evaluation sessions, relevant shots may be discovered, and the ground truth updated	for repeatable evaluation, a fixed ground truth set is desirable
although real world tasks may involve time pressure, it would be best to measure the time until the task is solved	time limits are needed in evaluation sessions for practical reasons

KIS and AVS, two task categories

Tasks at VBS – Known-item Search

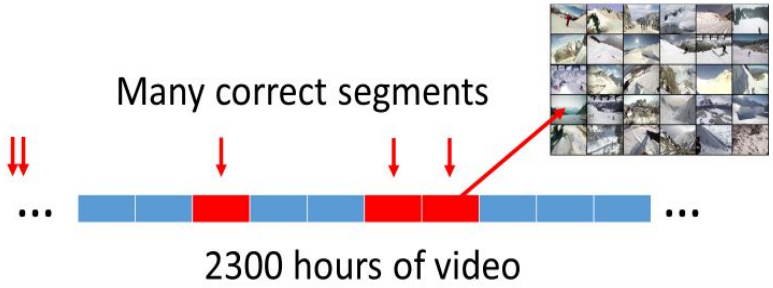
Target scene is known to the searcher, but no knowledge about location/position in dataset

Visual or textual input for users

Tasks at VBS – Ad-hoc Video Search

Find **many scenes** for a specific content class/topic

For example: "outdoor shots with snow or ice conditions"



How many task categories can be designed and tested?

Task Selection (KIS @ VBS)

- **Known duplicates:**
  - List of known (partial) duplicates from matching metadata and file size, content-based matches
- **Uniqueness inside same and similar content:**
  - Ensure unambiguous target
  - May be applied to sequence of short shots rather than single shot
- **Complexity of segment:**
  - Duration of roughly 20s
  - Limited number of shots
  - In fast-paced content, 20s may already be too long (too many shots)
- **Describability:**
  - Textual KIS requires segments that can be described with limited amount of text (less shots, salient location or objects, etc.)
  - Target must be described uniquely with the provided textual information

VBS KIS Task Selection - Examples

- **KIS Visual (video 02630, frame 750-1250)**
  - few shots from a wakeboarding scene - hard to describe as text, but unique sequence
- **KIS Textual (video 36496, frame 0-598)**
  - @0 sec: “Shots of a factory hall from above. Workers transporting gravel with wheelbarrows. Other workers putting steel bars in place.”
  - @100 sec: “The hall has Cooperativa Agraria written in red letters on the roof.”
  - @200 sec: “There are 1950s style American cars and trucks visible in one shot.”

Task Presentation

- **Visual KIS**
  - Originally, we simply repeated the 20s clip until the task ended
  - In order to simulate a ***fading memory***, we incrementally blurred the clip (2018-2019)
  - At VBS2020 we experimented with a blurred and color-less presentation from the start, highlighting only a ***salient object***
- **more options**
  - Play query ***once***: one chance to memorize, but not chance to check possibly relevant shot against query — like in real life, but there searcher has information need in mind, while VBS participant needs to learn it from the query
  - Depends on available technology: taking pictures of query to start search was still difficult in early years
  - Present query long enough ***before search***: closer to real situation, but introduces even larger dependency on human factors (visual memory abilities)

Task Presentation

- **Textual KIS**
  - Incrementally reveal details to simulate that searcher has a conversation with ***a domain expert, who can be asked questions or remembers more details***
  - This makes this kind of tasks much harder
  - First hint alone designed to discriminate target shot, but may be difficult to decide

0s: *"Shots of a factory hall from above. Workers transporting gravel with wheelbarrows. Other workers putting steel bars in place."*
100s: *"The hall has Cooperativa Agraria written in red letters on the roof."*
200s: *"There are 1950s style American cars and trucks visible in one shot."*

Retrieval queries at VBS

Ad-hoc Video Search (AVS)

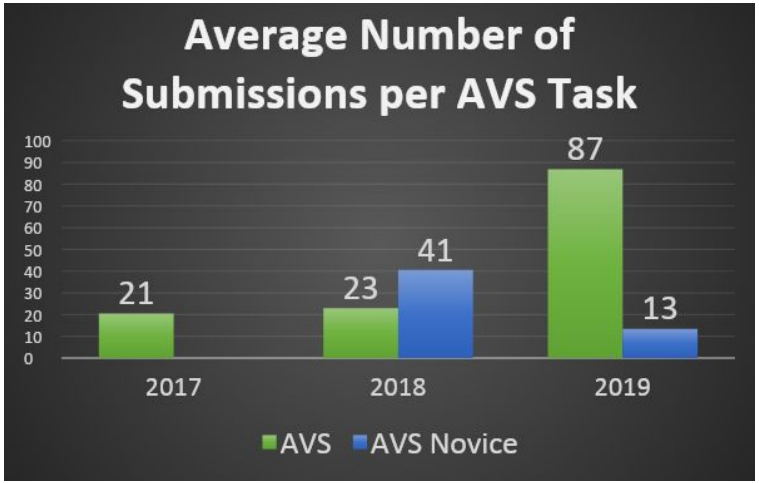
- Since 2017: in collaboration with TRECVID AVS
- We want to find **many scenes** for a specific content class/topic
- For example:
- ”an adult person running in a city street”
- ”a chef or cook in a kitchen”
- ”outdoor shots with snow or ice conditions”

Teams should solve KIS and AVS tasks as quickly and accurately as possible (max 5 mins per topic/7 mins for KIS textual)

- Awad, G., Butt, A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., ... & Kraaij, W. (2018, November). Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search.

AVS Tasks

- **Quite challenging to evaluate**
  - No complete ground truth (neither we nor TRECVID)
  - Our solution: use **judges** to perform live evaluation of submissions for which we do not have G.T. yet



AVS task definition dilemma

- **Find a compromise in terms of the number of expected results**
- **broader == larger result set**
  - may be closer to some real content needs
  - chance for more fine grained performance assessments due to large possible number of results
  - compatibility with TRECVID queries
  - too many results to judge
    - having result live during competition is part of the VBS experience
- **more specific == smaller result set**
  - queries involve conditions and negations - realistic challenge
    - e.g. shots of people kissing who are not bride and groom
  - content-wise the some queries appear as “constructed”
  - sometimes few samples that occur in the same video (i.e. teams find nothing or all)

Evaluation of AVS Tasks at VBS

- **1,848 shots judged live in 2017 (2018: 2,780 shots)**
  - About 40% of submitted shots were not in TRECVID G.T.
- **Verification experiment**
  - 1,383 shots were judged again later
  - Judgements were diverging for 23% of the shots, in 88% of those cases the live judgement was “incorrect”
- **Judges seem to make incorrect decisions when in doubt**
  - But: while their decisions are biased, still same conditions for all teams in the room
- **Needed to set up clear rules for live judges**
  - Like used by NIST for TRECVID annotations

Judge 1:
false

Judge 2:
true

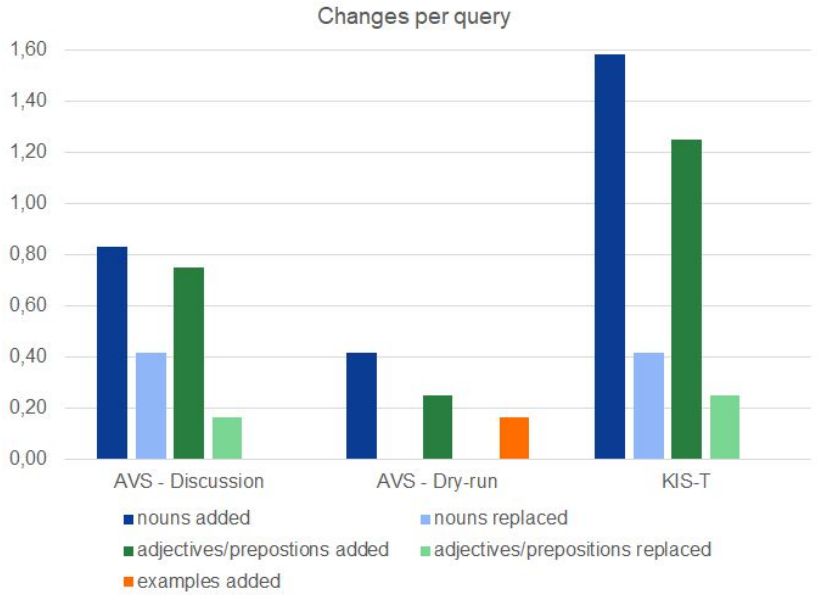
same video

Judge 1:
true

Judge **1**:
false

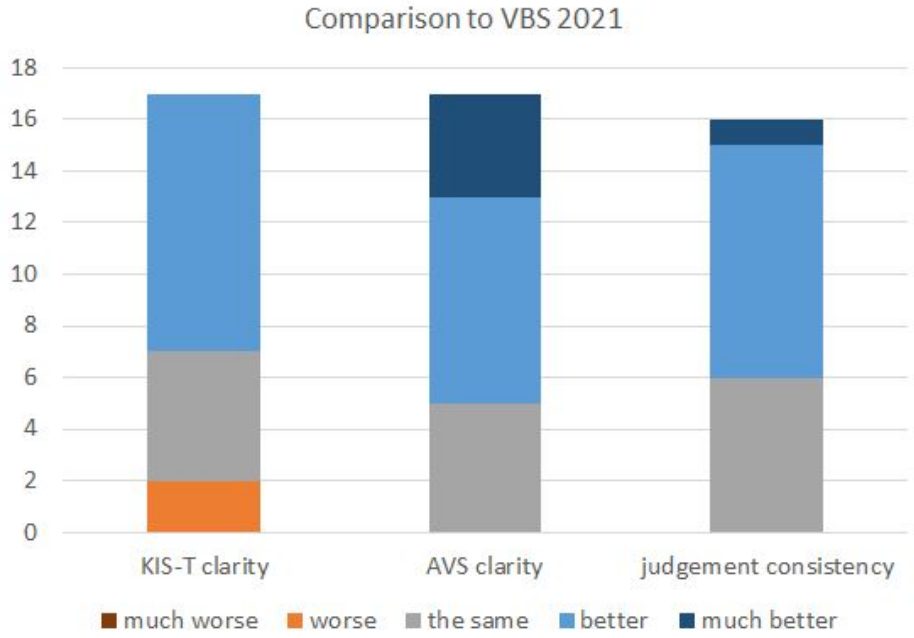
Judge Briefing

- **Up to 2020, judges informally discussed open issues coming up**
- **Online setting in 2021**
  - meeting in advance of competition discussing queries
  - did not result in sufficiently consistent judgements
- **Briefing in 2022**
  - meeting to discuss and refine AVS and KIS-T queries “on paper”
  - dry-run: judges try solving AVS queries (not aiming at high recall, but until a sufficiently divers result set is found)
  - queries changes significantly



Judge Briefing

- **Briefing in 2022 - preliminary evaluation**
  - survey among VBS participants
  - n=20, 17 also participated in 2021
  - comparison of task clarity
  - comparison of judgement consistency



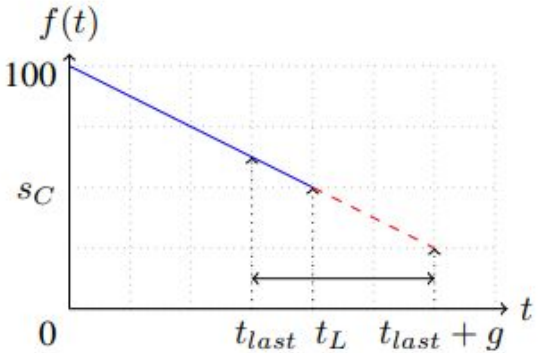
VBS Scoring

- **General goals**
  - Reward for (1) **solving a task** and for (2) **being fast**
  - Fair scoring and penalty for wrong submissions
- **Known-Item Search**

$$f^i_{KIS}(t, ws) = \lceil \max(0, s_C + (100 - s_C) \cdot f^i_{TS}(t) - f_{WSP}(ws)) \rceil$$

$$f^i_{TS}(t) = (t^i_L - t)/t^i_L$$

$$f_{WSP}(ws) = 10 \cdot ws$$



- s_C is time-independent reward for solving a task *i* (e.g., 50)
- f_{TS} is a linearly decreasing function, based on search time *t*
- *g* is a guarantee between the last accepted correct submission and the first potential late correct submission (e.g., 30s) – i.e. the time limit is extended by *g*

Visual KIS: 5 min
Textual KIS: 7 min

- J. Lokoč, W. Bailer, K. Schoeffmann, B. Muenzer and G. Awad, "On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015–2017," in *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3361-3376, Dec. 2018.

VBS Scoring

- **Ad-hoc Video Search**
  - Scoring based on **Precision** and **Recall** according to
    - **correct and incorrect submissions of the team (*C* and *I*)**
    - **pool of correct shot submissions of all teams for the task (*P*)**
    - **quantization function q that merges temporally close correct shots (into *ranges*; since VBS2018 ranges are fixed static non-overlapping segments of 180s duration)**
  - We mitigate impact of incorrect submissions to reduce penalty in case of ambiguous topic descriptions

$$f_{AVS}(C, I, P) = \left\lceil \frac{100 \cdot |C|}{|C| + |I|/2} \cdot \frac{|q(C)|}{|q(P)|} \right\rceil$$

- Jakub Lokoč, Gregor Kovalčík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. 2019. Interactive Search or Sequential Browsing? A Detailed Analysis of the Video Browser Showdown 2018. ACM Trans. Multimedia Comput. Commun. Appl. 15, 1, Article 29 (February 2019), 18 pages. DOI: https://doi.org/10.1145/3295663

VBS Scoring

- **Final score for a team *j* is the average score over all five categories, normalized by the corresponding maximum of each category/session *c***
  - Visual KIS expert
  - Textual KIS expert
  - Visual KIS novice
  - AVS expert
  - AVS novice

$$\left\lceil \frac{1}{5} \cdot \sum_{c=1}^{5} \frac{100 \cdot score^{c}_{team_j}}{\max_{j=1..k}(score^{c}_{team_j})} \right\rceil$$

What are we measuring?

- **Apart from the task design, the choice of the metric shapes the real-world information need being modelled**
- **Which aspects are important? How are they (implicitly) encourage/discouraged by the metric?**
- **KIS**
  - as there is one correct results, this aspect is simple
  - how much focus on submission time?
  - tolerance for false submissions?
  - weight between comprehensive results and speed

What are we measuring?

- **AVS**
  - unlikely to have complete result set returned by the task
- **aspects included in metric**
  - comprehensiveness of results (how many of the known relevant shots)
  - correctness: via penalty for false submissions
  - speed: if tasks were completely solvable, we wouldn’t measure it with the current metric, but for most tasks most teams do not finish in the working time
  - diversity: via ranges within content (may not always be semantically diverse) - results from different videos preferred

What are we measuring?

- **issues with AVS metric**
  - probably favours many submissions: teams submitting many mostly correct segments seem to do better than teams submitting cherry-picked correct ones
  - measuring diversity:
    - **ranges have unwanted side-effects: chains of overlapping segments scored differently than scattered segments**
    - **reliable segmentation is not easy to obtain: semantically meaningful unit may not be independent of query**
  - as with all metrics using pooling, repeatability is not fully guaranteed, when new relevant shots are found in later runs
- **redesign**
  - metric redesign may involve task redesign, i.e. to limit amount of relevant segments

Where is deep learning helpful?

(And where not?)

Summary of relevant DL achievements

Since 2012, deep learning methods win various benchmark challenges

For interactive video search, the following approaches are highly relevant

1. Image/video classification, object detection, semantic segmentation, automatic annotation, event detection - all based on deep neural networks
2. For text-image search, joint embedding approaches can use very large train datasets (e.g., CLIP used 400M text-image pairs)
3. Similarity search methods rely on deep learning as well
4. Video structure analysis (e.g., shot detection)

Summary of relevant DL achievements

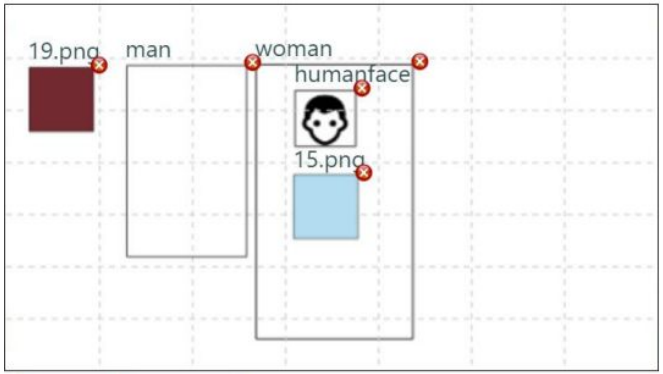
Since 2012, deep learning methods win various benchmark challenges

For interactive video search, the following approaches are highly relevant

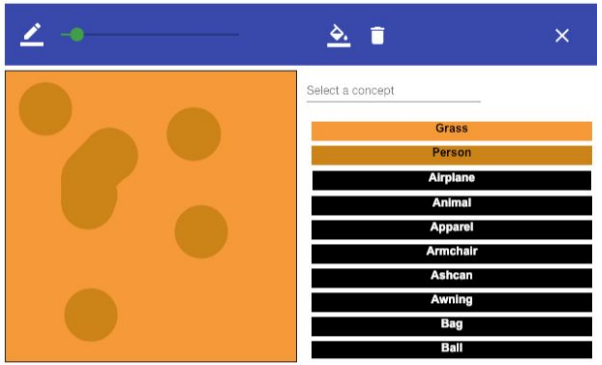
1. Image/video classification, object detection, semantic segmentation, automatic annotation, event detection - all based on deep neural networks
2. For text-image search, joint embedding approaches can use very large train datasets (e.g., CLIP used 400M text-image pairs)
3. Similarity search methods rely on deep learning as well
4. Video structure analysis (e.g., shot detection)

How can the achievements help at VBS?

Query specification allows to enter class keywords, object locations, semantic sketch, text query, and event class in GUI



VISIONE system



vitrivr system

https://arxiv.org/pdf/2008.02749.pdf

https://vitrivr.org/

How can the achievements help at VBS?

Query specification allows to enter class keywords, object locations, semantic sketch, text query, and event class in GUI

Where such querying does not lead to success?

- Dataset contains too much items of the query class
- Despite state-of-the-art performance, DL methods can still fail on “wild data”
- Users find it difficult to identify proper query (e.g., out of 10000 classes)
- With rich GUI, users can focus on less effective model for a given task

Summary of relevant DL achievements

Since 2012, deep learning methods win various benchmark challenges

For interactive video search, the following approaches are highly relevant

1. Image/video classification, object detection, semantic segmentation, automatic annotation, event detection - all based on deep neural networks
2. For text-image search, joint embedding approaches can use very large train datasets (e.g., CLIP used 400M text-image pairs)
3. Similarity search methods rely on deep learning as well
4. Video structure analysis (e.g., shot detection)

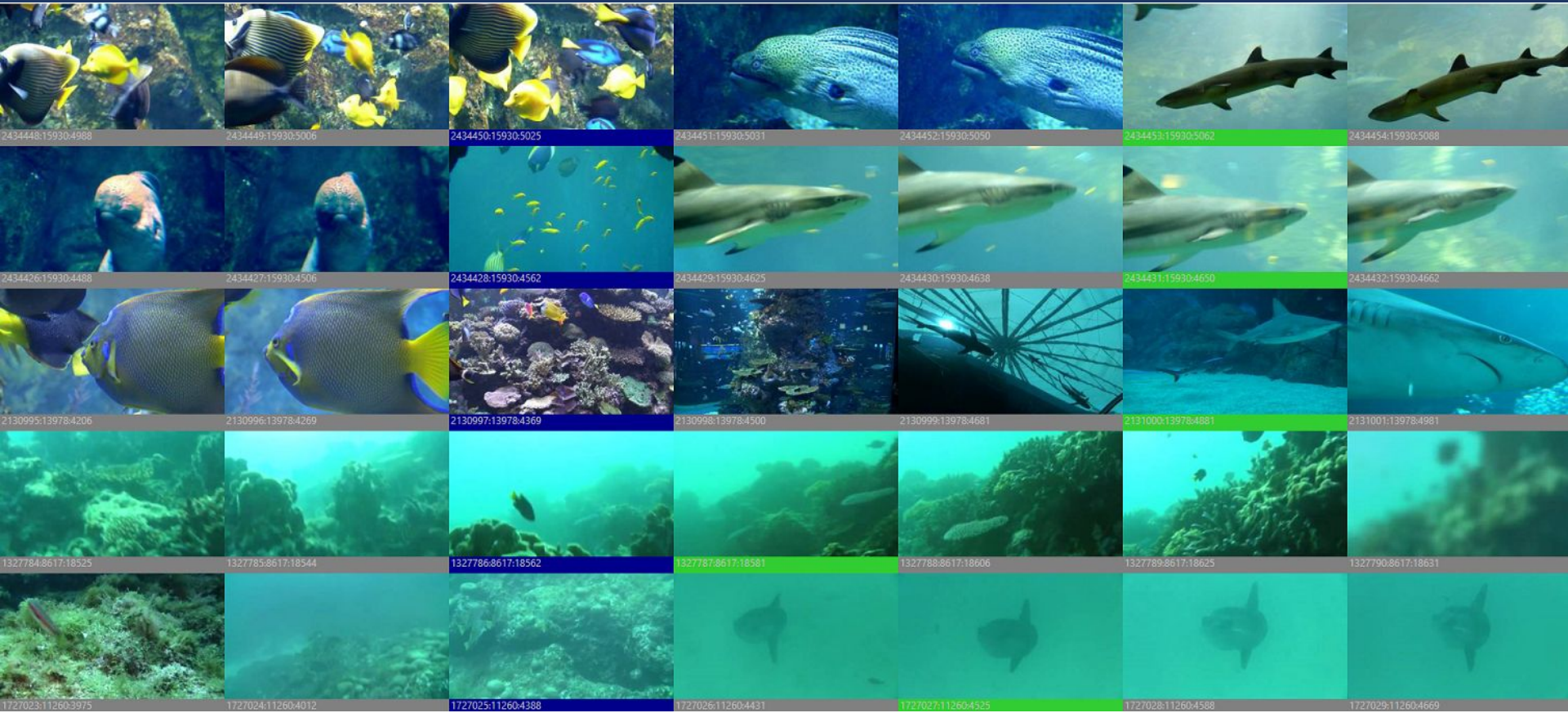
How can the achievements help at VBS?

Joint embedding methods allow convenient free-form text query specification, allowing specification of more details about a searched scene. **Currently the most promising approach at VBS and LSC campaigns.**

Let’s try some queries using CLIP model and V3C collection, where text and images are mapped to a joint vector space with cosine similarity

query: “green bird sitting on branch”

query: “coral reef with a yellow fish > shark”



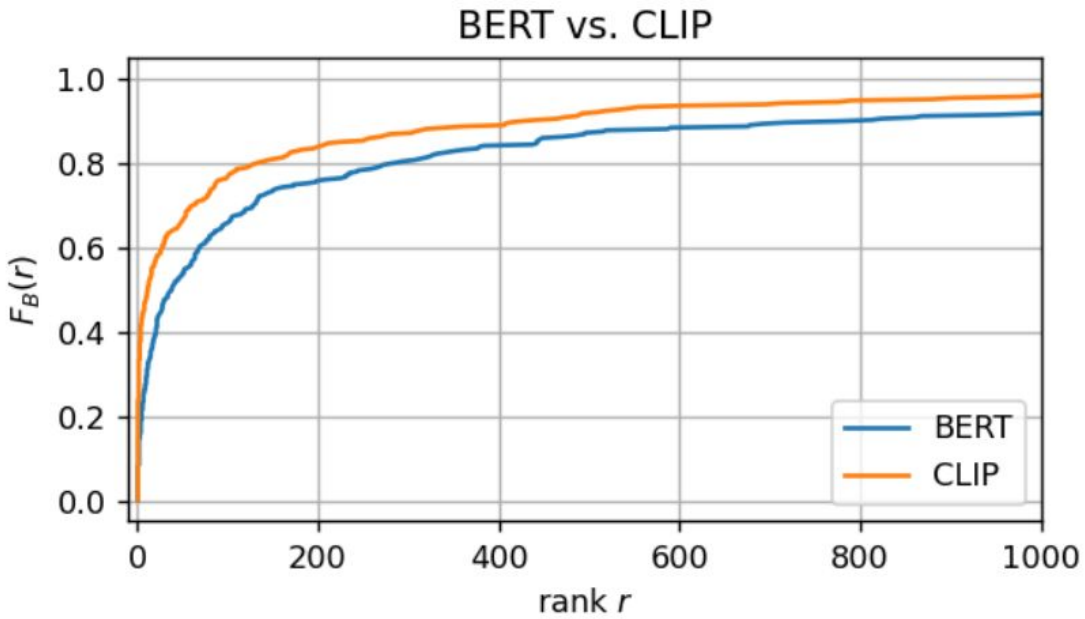
How can the achievements help at VBS?

Joint embedding methods allow convenient free-form text query specification, allowing specification of more details about a searched scene. **Currently the most promising approach at VBS and LSC campaigns.**

Where joint embedding based search still does not lead to success?

- Users may still find it difficult to identify proper text query (gym or sport hall?)
- User is not familiar with supported language or domain vocabulary (fish?)

How can the achievements help at VBS?



$$F_B(r) = \frac{|\{r_i : r_i \in Ranks, r_i \leq r\}|}{|Ranks|}$$

Ranks obtained for 327 text-target pairs, where for each pair the text query is used to rank 20K image dataset and find the target image

J. Lokoč, T. Souček: How Many Neighbours for Known-item Search?

Summary of relevant DL achievements

Since 2012, deep learning methods win various benchmark challenges

For interactive video search, the following approaches are highly relevant

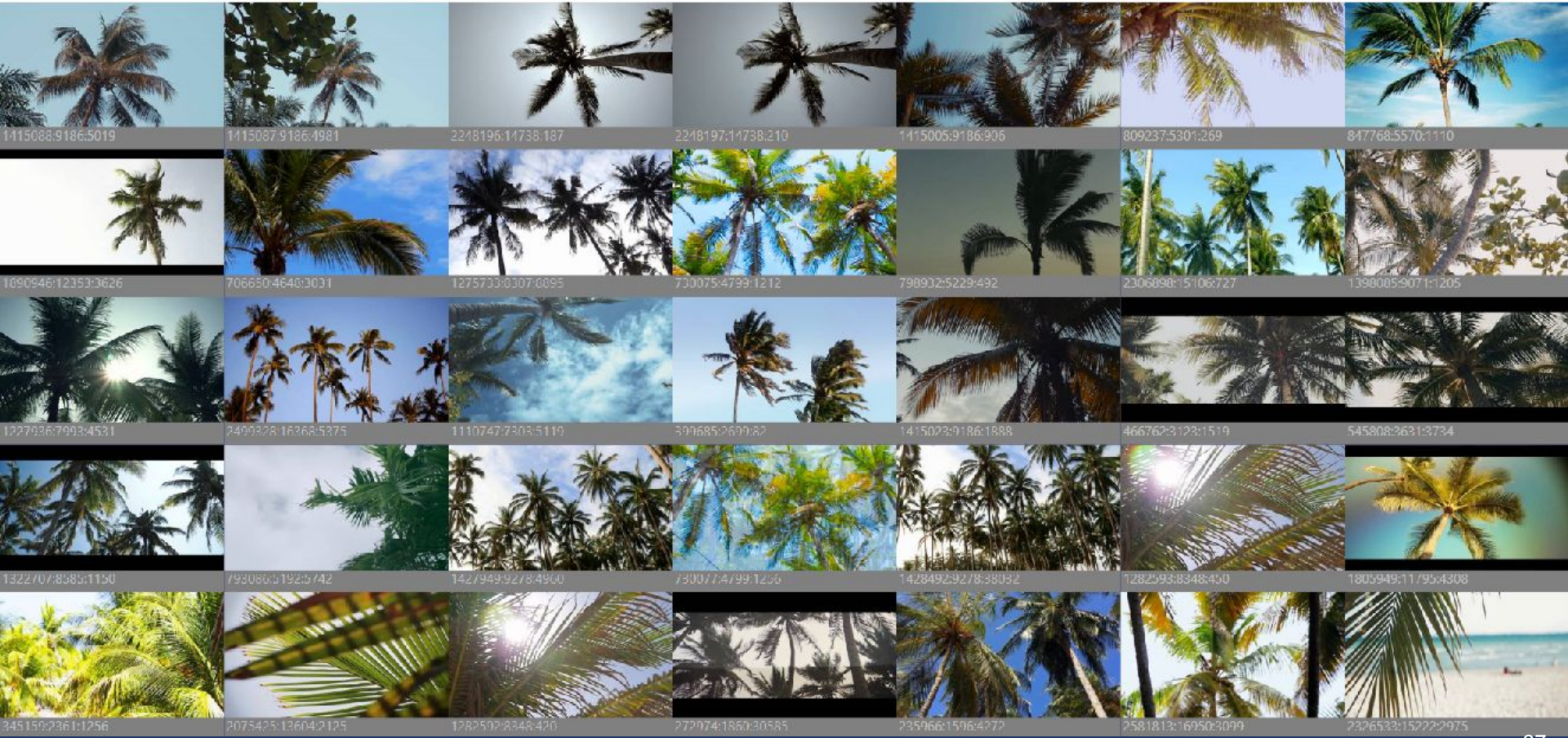
1. Image/video classification, object detection, semantic segmentation, automatic annotation, event detection - all based on deep neural networks
2. For text-image search, joint embedding approaches can use very large train datasets (e.g., CLIP used 400M text-image pairs)
3. Similarity search methods rely on deep learning as well
4. Video structure analysis (e.g., shot detection)

How can the achievements help at VBS?

Once users observe results, refinement based on selected example images can help to get closer to target segment frames. Similarity model can be used for kNN search or for a relevance feedback score update rule

Let’s try an example image of a palm tree, again CLIP model (though there exist better models for similarity search)

How can the achievements help at VBS?



How can the achievements help at VBS?

Once users observe results, refinement based on selected example images can help to get closer to target segment frames. Similarity model can be used for kNN search or for a relevance feedback score update rule

Where similarity search does not lead to success?

- User has a different idea of visual appearance of the target scene (textual KIS)
- User is not familiar with the similarity model (what does it mean similar?)

Summary of relevant DL achievements

Since 2012, deep learning methods win various benchmark challenges

For interactive video search, the following approaches are highly relevant

1. Image/video classification, object detection, semantic segmentation, automatic annotation, event detection - all based on deep neural networks
2. For text-image search, joint embedding approaches can use very large train datasets (e.g., CLIP used 400M text-image pairs)
3. Similarity search methods rely on deep learning as well
4. Video structure analysis (e.g., shot detection)

How can the achievements help at VBS?

4. Video structure analysis (e.g., shot detection)

With shot boundaries available, it is possible to

- Design task categories with shot based search unit (e.g., no car in the shot)
- Use shot boundaries to design new video search approaches

There are open-source tools (e.g., TransNet) for very fast and also effective detection of common shot transitions using DCNNs

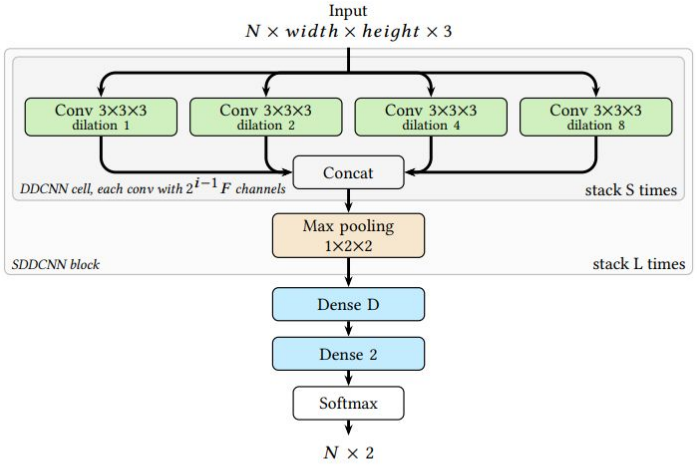
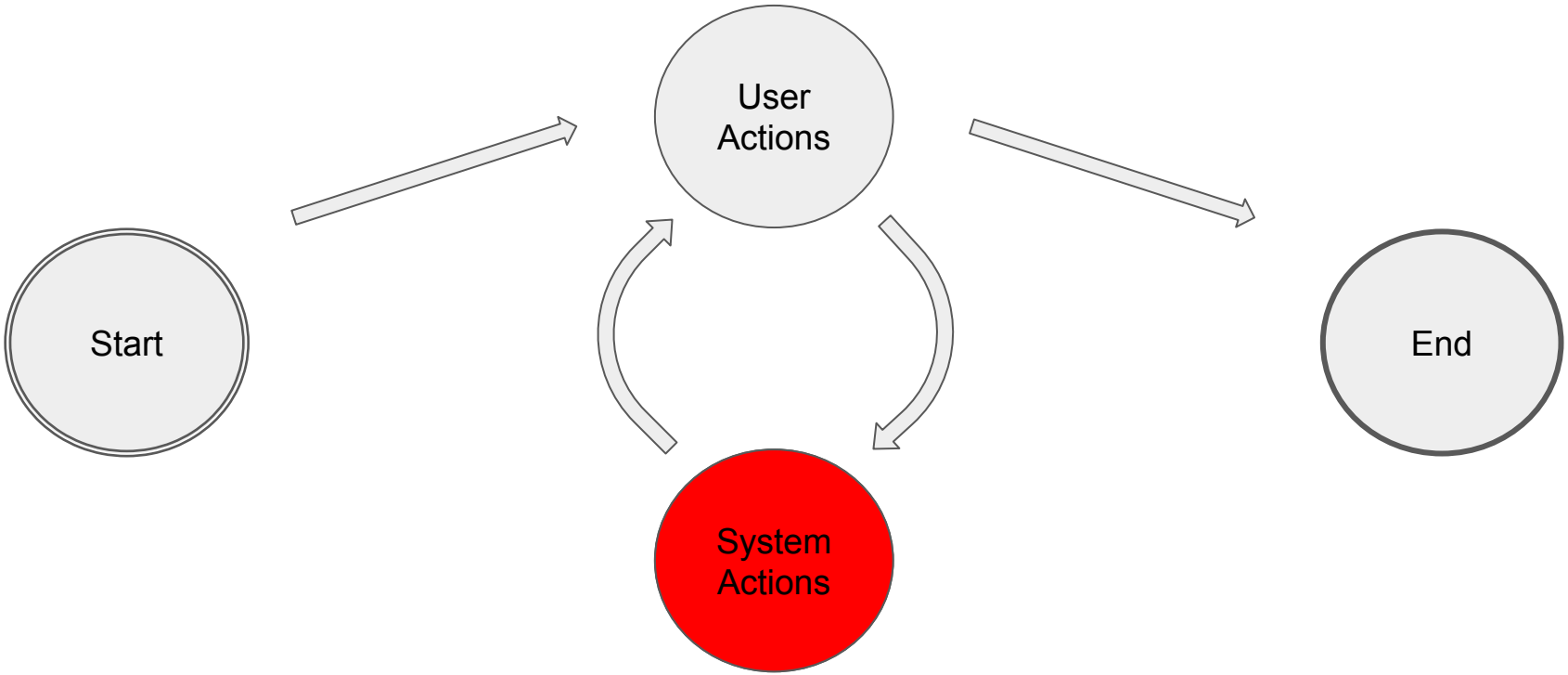


Figure 1: TransNet shot boundary detection network architecture for $S = 1$ and $L = 1$. Note that N represents length of video sequence, not batch size. In our case $N = 100$.

T. Souček et al: TransNet: A deep network for fast detection of common shot transitions

Evaluating Implementation Choices

Interactive System – Simplified View



Evaluating the Engine: Options

Focus on parameters and performance	Focus on usability and experience
During system development	The “final word”
Cheap and fast	Expensive and time-consuming
Reproducible	Non-reproducible
Not representative of interactions	Based on interactions

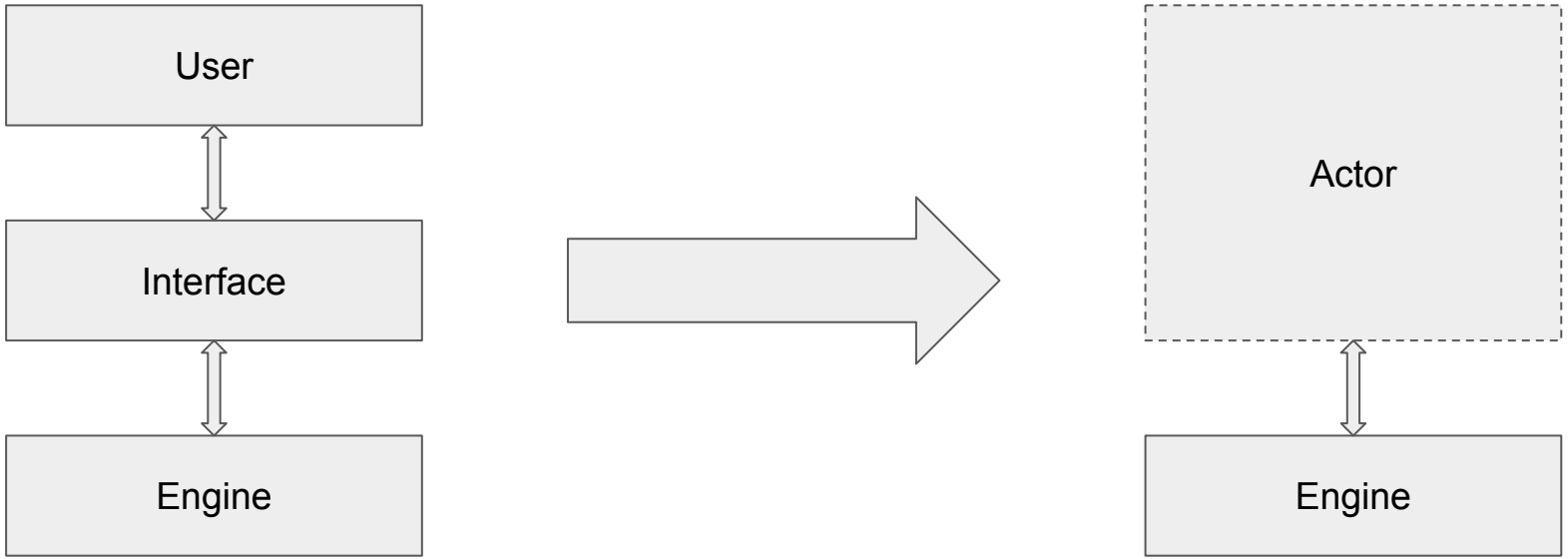
Benchmarks **User Studies**

Evaluating the Engine: Options

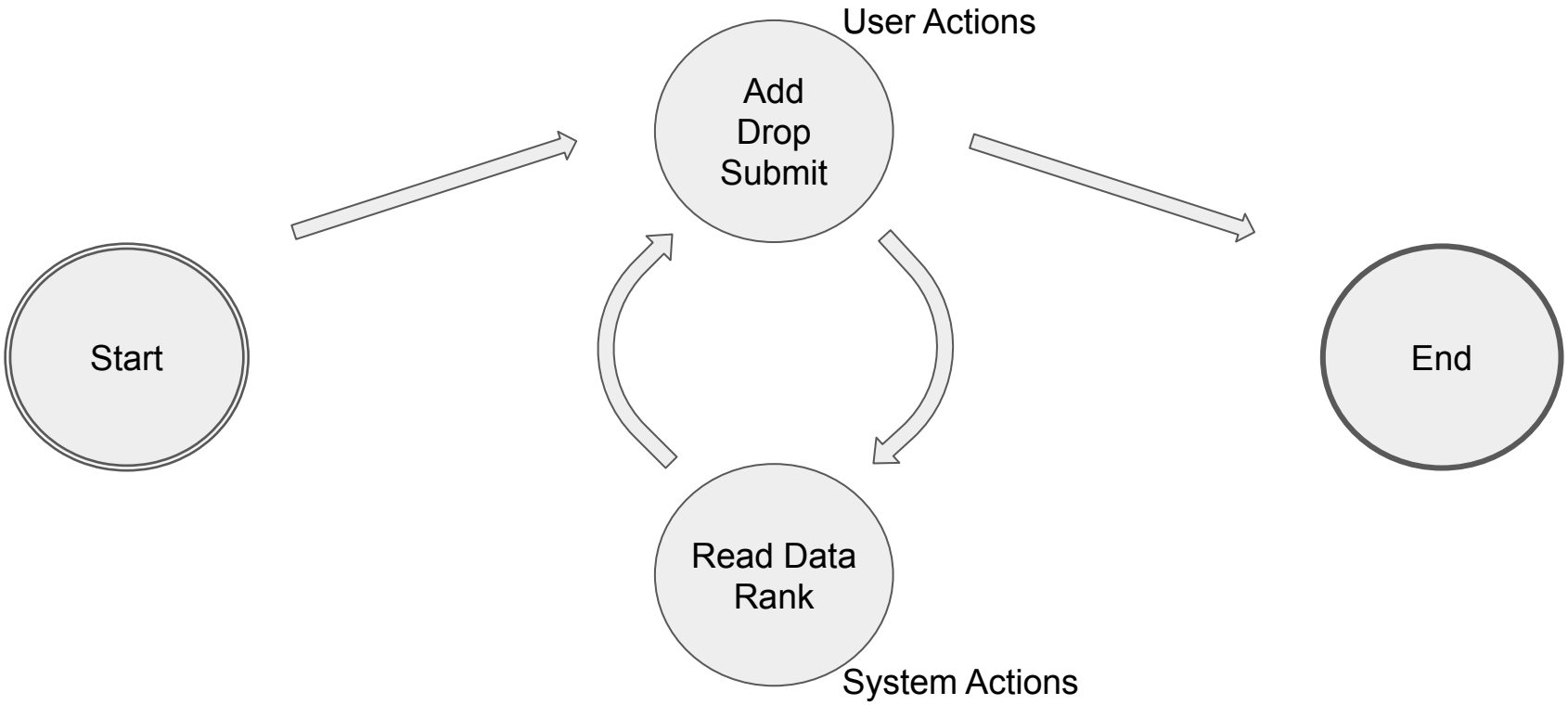
Focus on parameters and performance	Focus on usability and experience
During system development	The “final word”
Cheap and fast	Expensive and time-consuming
Reproducible	Non-reproducible
Not representative of interactions	**Based on interactions**

Benchmarks — **Actors = Artificial Users** — User Studies

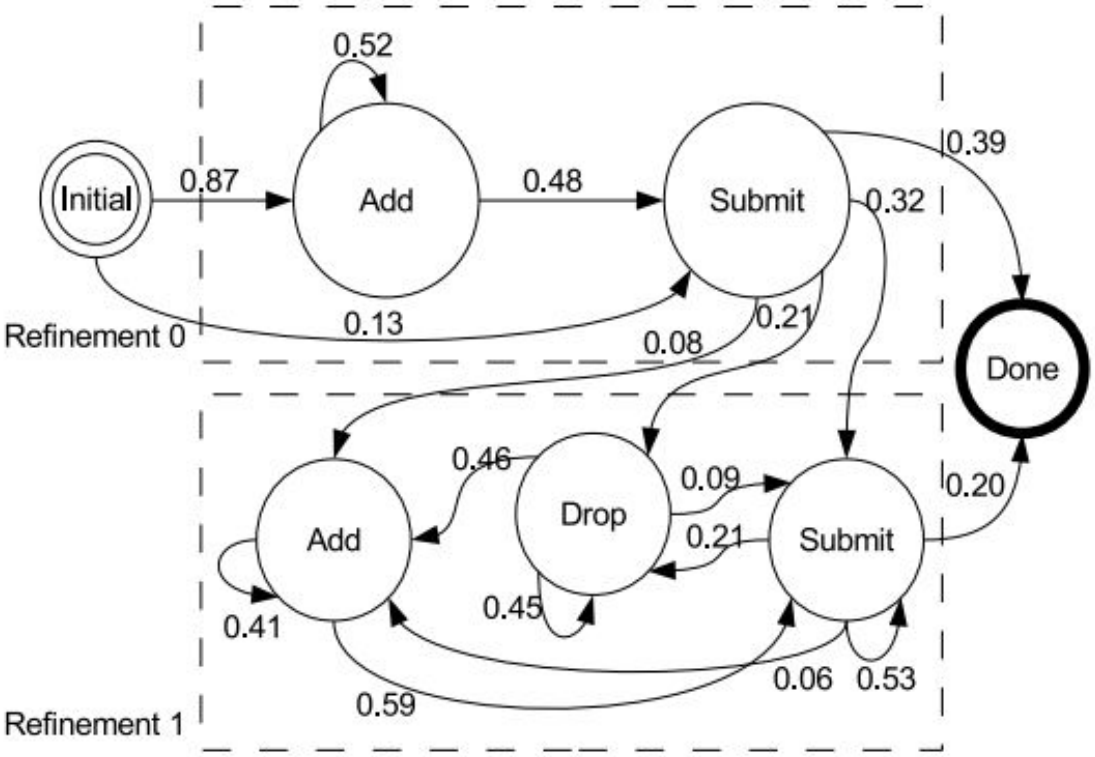
Focus on the Engine: Simulated Users



Interaction Example: Web Search

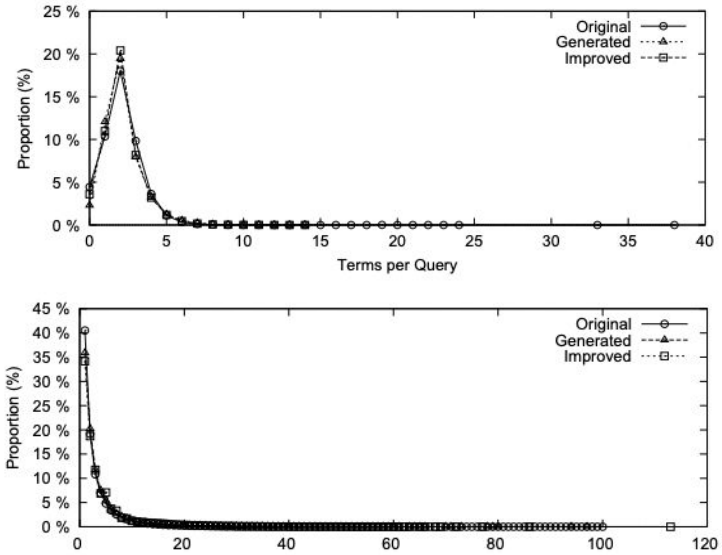


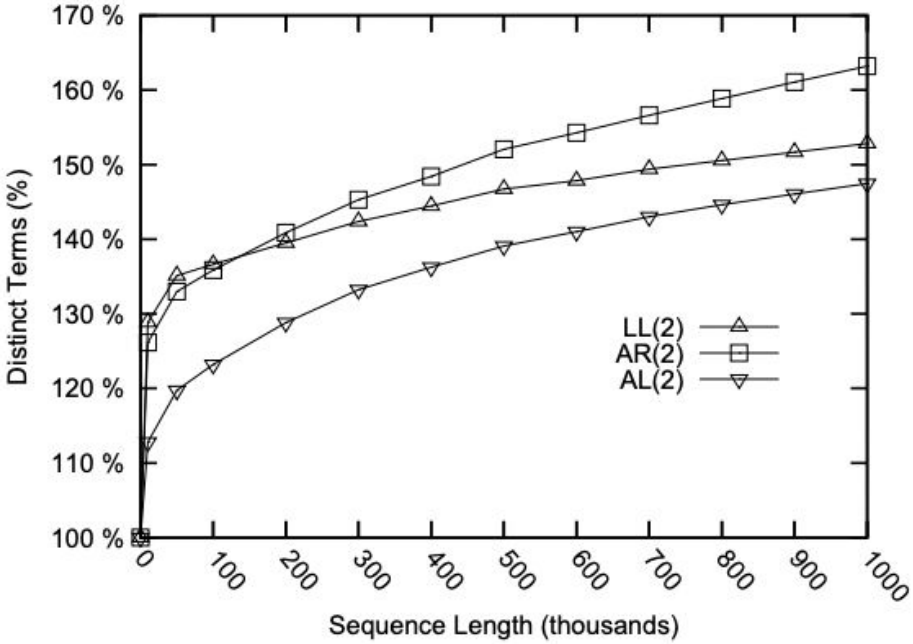
Web Search Actors: Excite Log Analysis



Sigurþórsdóttir: Towards Automatic Generation of Realistic Web Query Sequences. MSc thesis, 2011

Web Search Actors: Simulation Results

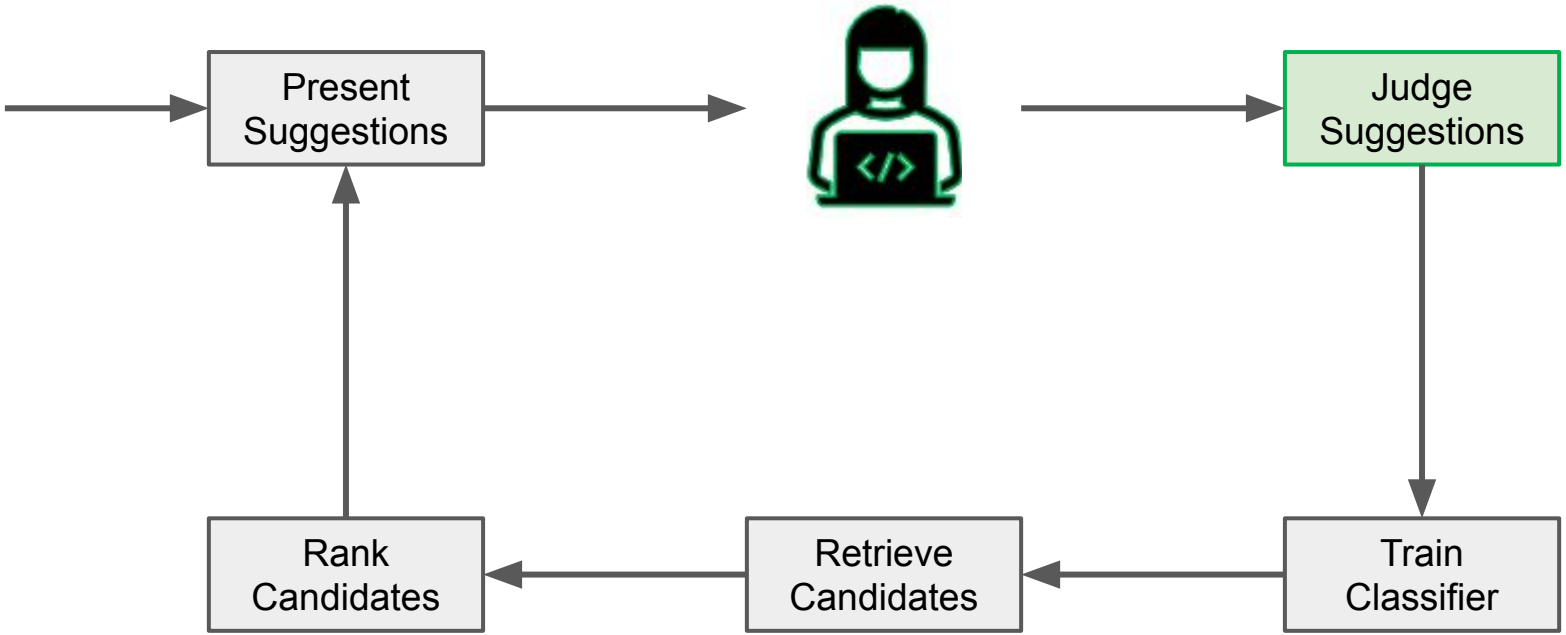




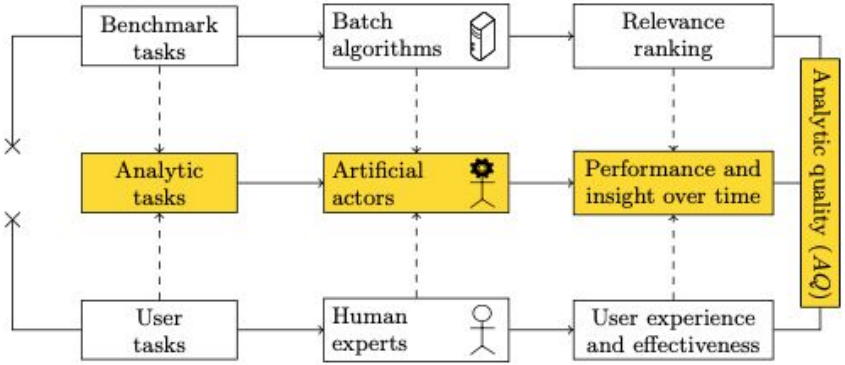
← Accurate for in-memory processing – Inaccurate for disk-based processing →

Sigurþórsdóttir: Towards Automatic Generation of Realistic Web Query Sequences. MSc thesis, 2011

User Relevance Feedback



Analytic Quality – Simulating Exploration

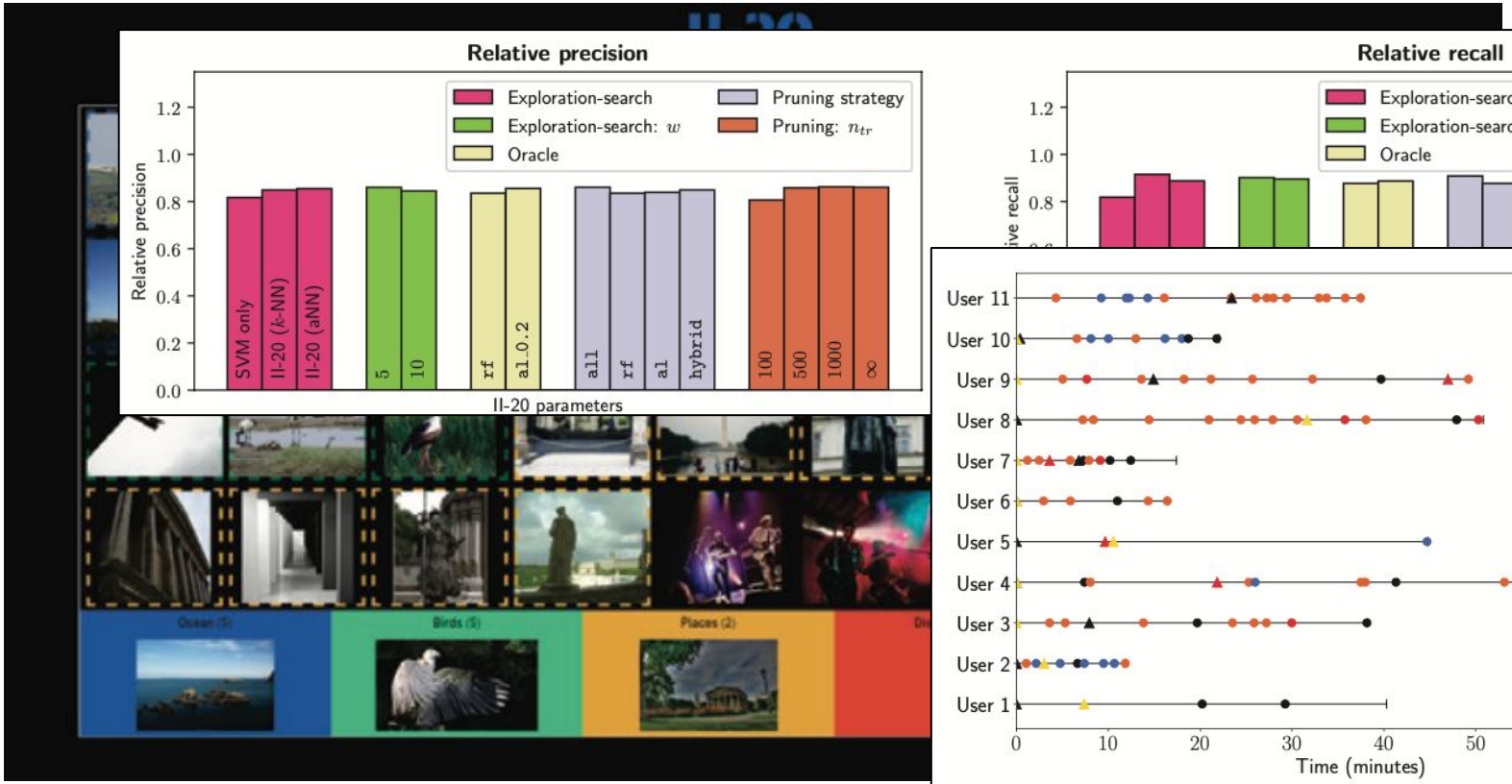


Actors

1. Initial categories
   - 7±2 groups of media items *{cat}, {dog, bone}*
   - Requires **ground truth** which is **independent** of evaluated architectures
2. Number of insight changes
   - low = expert, high = novice
3. Insight change times
   - long = focused user
4. Insight change actions
   - Add category *{cat}* → *{cat}, {dog}*
   - Remove category *{cat}, {dog}* → *{cat}*
   - Replace category *{cat}* → *{dog}*
   - Expand category *{dog}* → *{dog, bone}*
   - Reduce category *{dog, bone}* → *{dog}*
   - Change category *{dog, bone}* → *{dog, toy}*

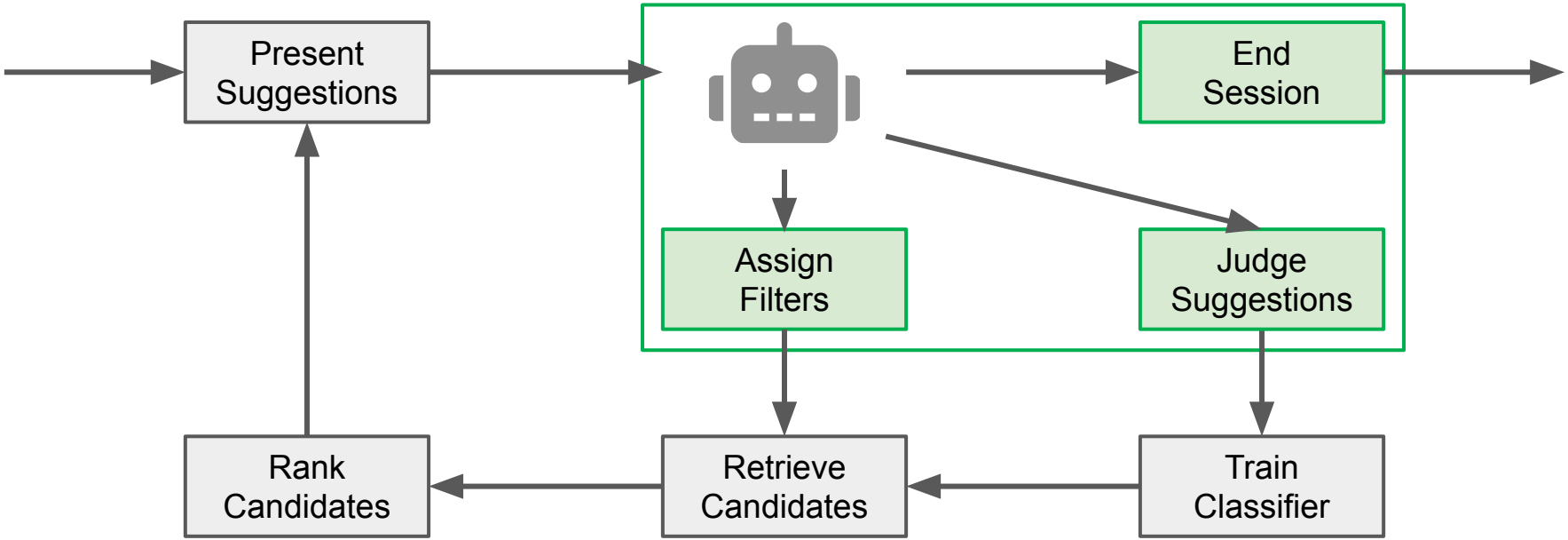
Zahálka et al.: Analytic Quality: Evaluation of Performance and Insight in Multimedia Collection Analysis. ACM MM 2015

Application of Analytic Quality: II-20



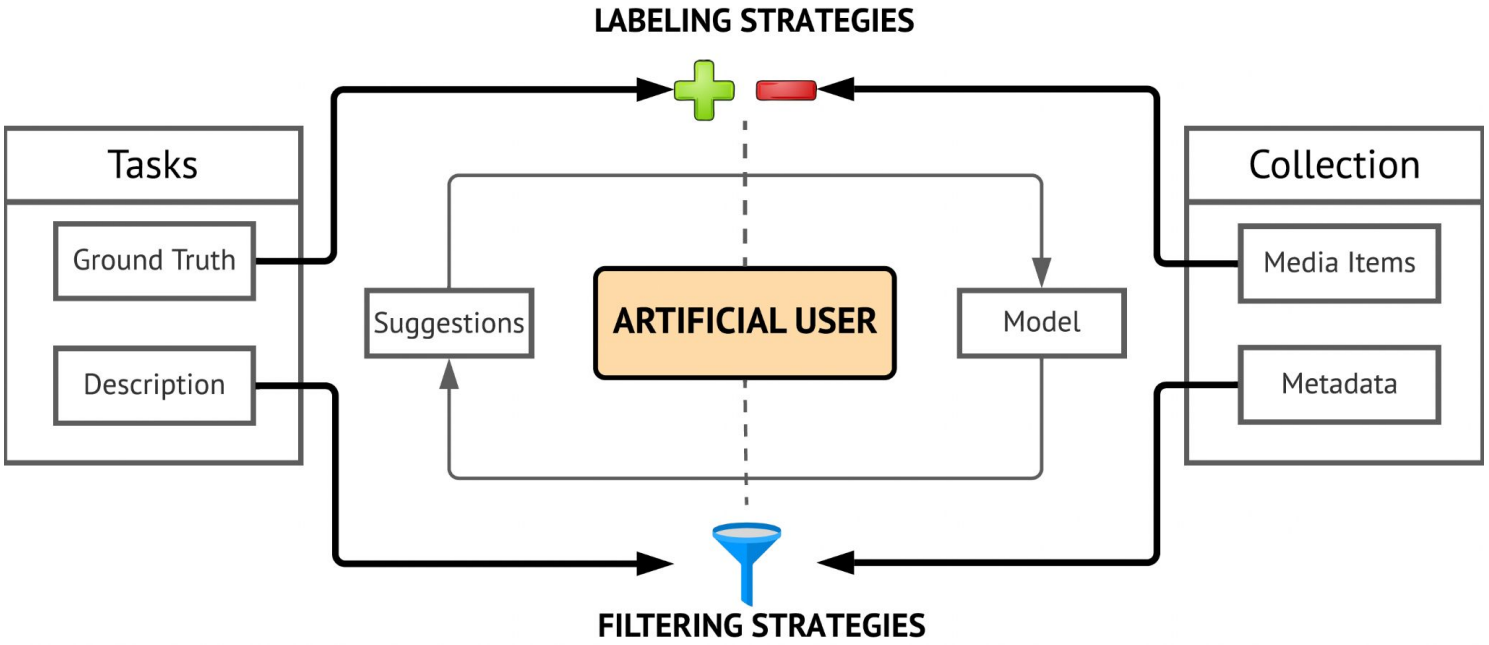
Zahálka et al.: II-20: Intelligent and pragmatic analytic categorization of image collections. IEEE TVCG 2021

Can We Simulate Interactive Retrieval Tasks?



Khan et al.: Impact of Interaction Strategies on User Relevance Feedback. ICMR 2021

Evaluation Protocol with Interaction Strategies



Khan et al.: Impact of Interaction Strategies on User Relevance Feedback. ICMR 2021

Labeling Strategies

Accumulative: Keep appending to positive and negative sets

1. Label *p* and *n* items in each round to the positive set *P* and negative set *N* (Add)
2. Allow user to replace items from the positive set *P* and negative set *N* (Replace)

Fixed: Strategies that use limited number of training examples

3. Limit positive set *P* and negative set *N* to *p* and *n* items (Both)
4. Limit only positive set *P* to *p* items (Positive)

Arbitrary Negatives: System labels negatives instead of user

5. Apply negatives from the suggestion set *S* (Local)
6. Apply negatives from the whole collection (Global)

Khan et al.: Impact of Interaction Strategies on User Relevance Feedback. ICMR 2021

Filtering Strategies

No Filter: Never choose to apply filters

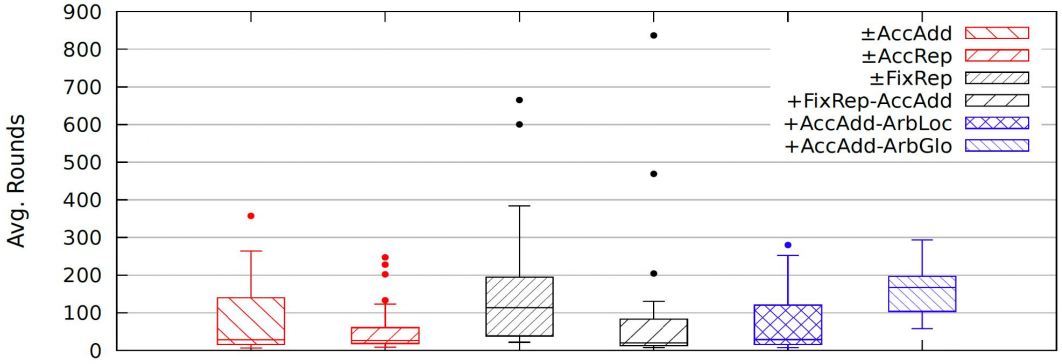
Novice: Apply based on terms in the task description

Expert: Infer filters from task description and correct wrong filters applied in earlier round

Data Author: Knows the collection in depth and was part of curating the metadata

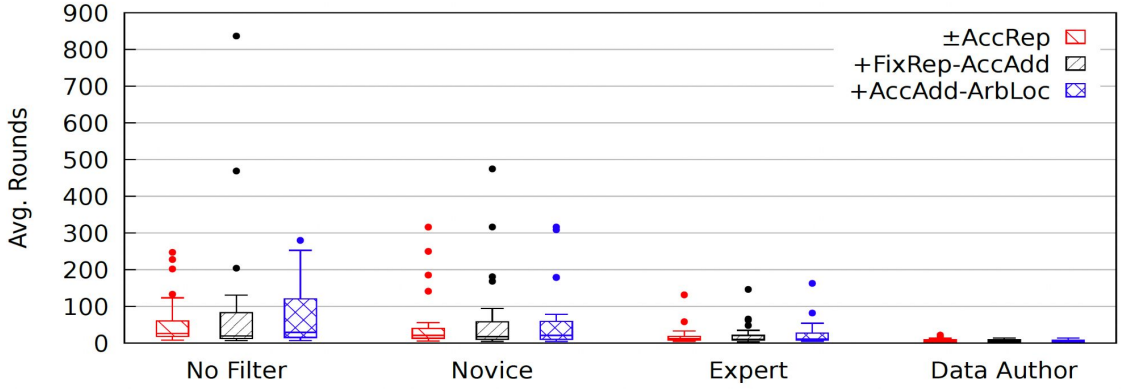
Khan et al.: Impact of Interaction Strategies on User Relevance Feedback. ICMR 2021

Dataset: Lifelog Search Challenge 2019 (LSC2019)



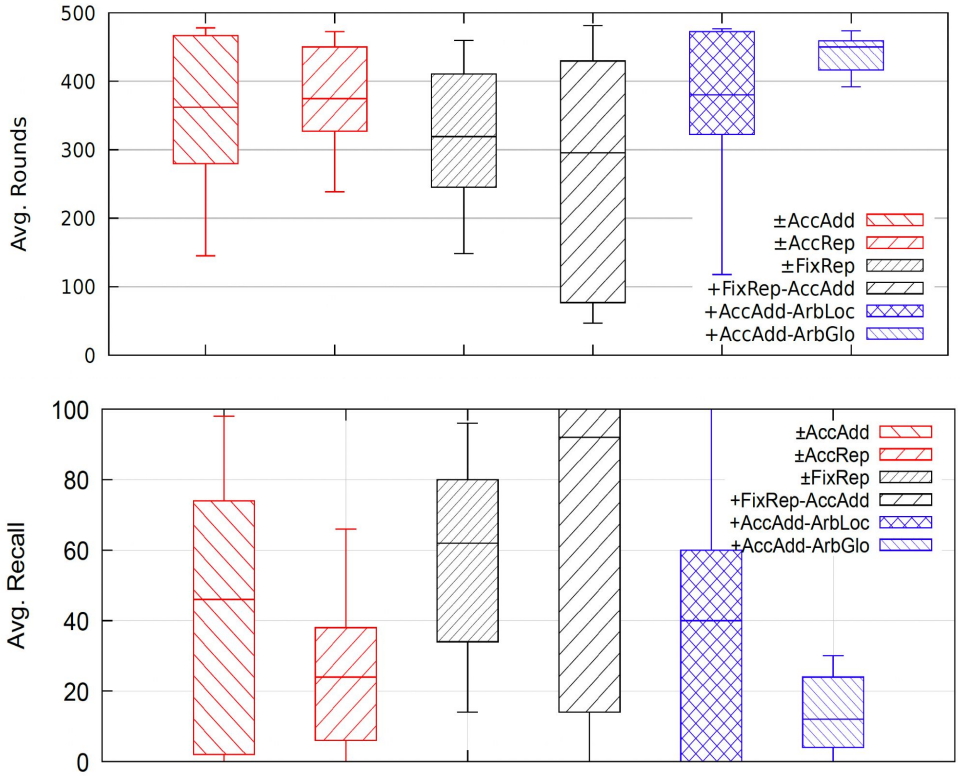
Accumulative (1: Add, 2: Replace)
Fixed (3: Both, 4: Positive)
Arbitrary Negatives (5: Local, 6: Global)

Accumulative (2: Replace)
Fixed (4: Positive)
Arbitrary Negatives (5: Local)



Khan et al.: Impact of Interaction Strategies on User Relevance Feedback. ICMR 2021

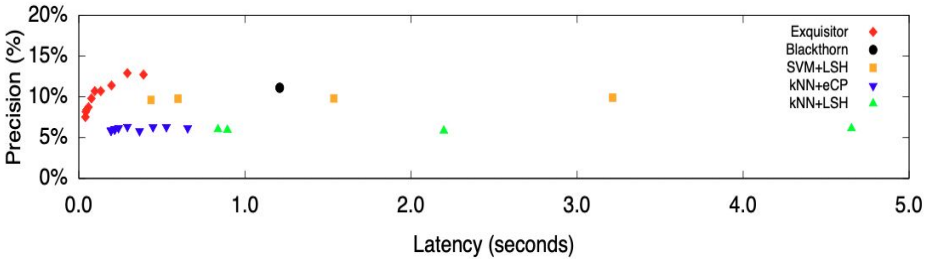
Dataset: Video Browser Showdown 2020 (VBS2020)

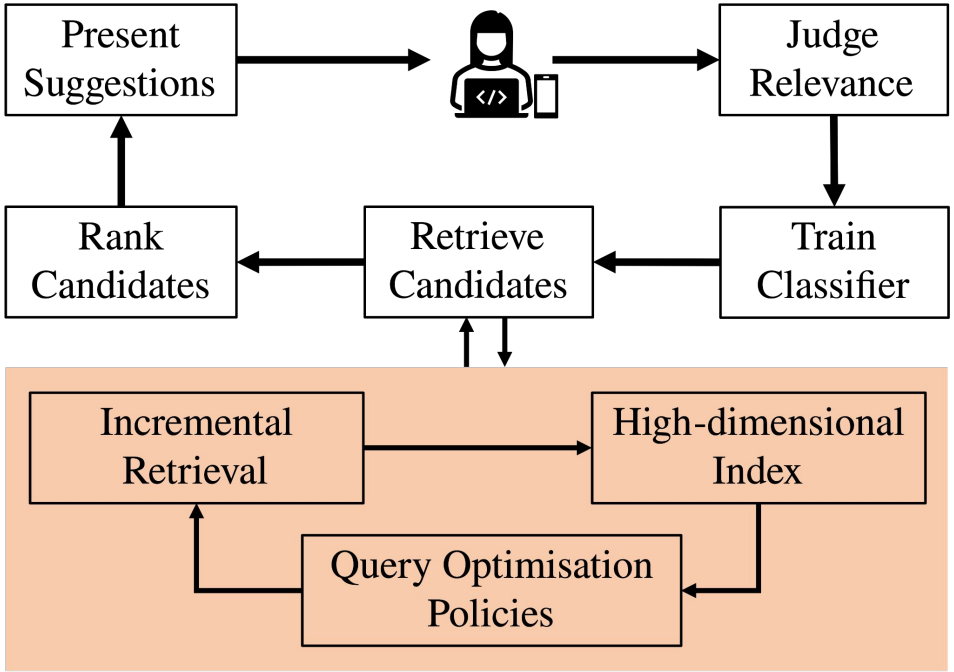


Accumulative (1: Add, 2: Replace)
Fixed (3: Both, 4: Positive)
Arbitrary Negatives (5: Local, 6: Global)

Khan et al.: Impact of Interaction Strategies on User Relevance Feedback. ICMR 2021

Application: Can We Find Needles in Haystacks?



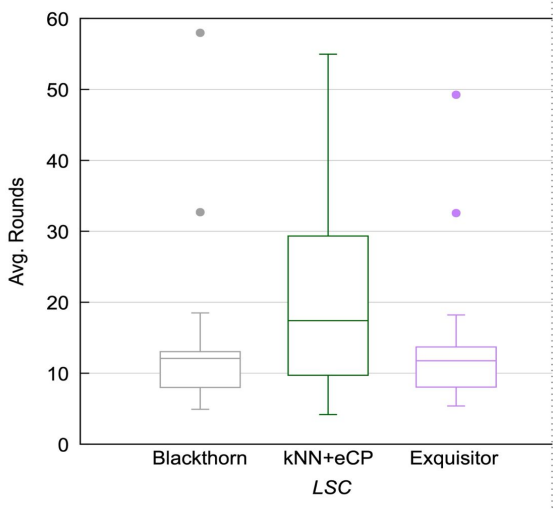


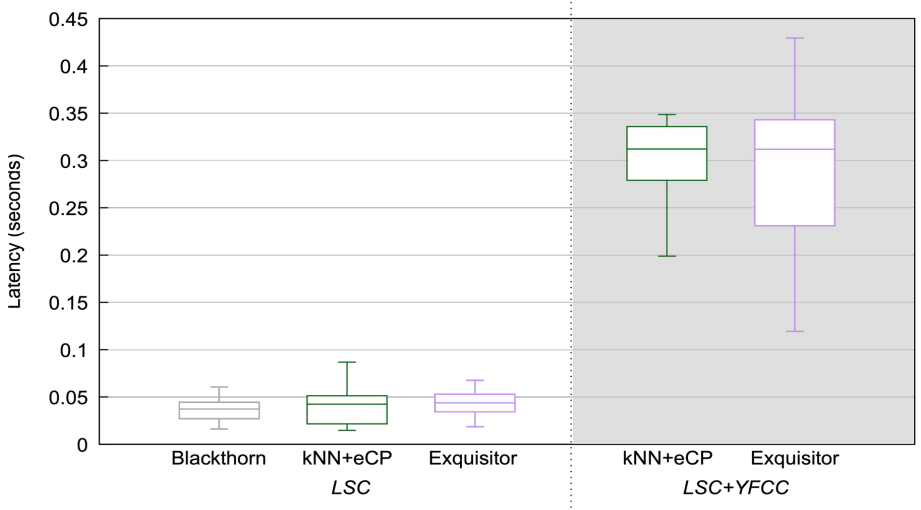
Khan et al: Interactive Learning for Multimedia at Large. ECIR 2020

LSC 2019 + YFCC 100M

Needle = Solving LSC tasks
Haystack = YFCC 100M (= 2500x LSC)

Quality is unaffected, latency same as YFCC alone

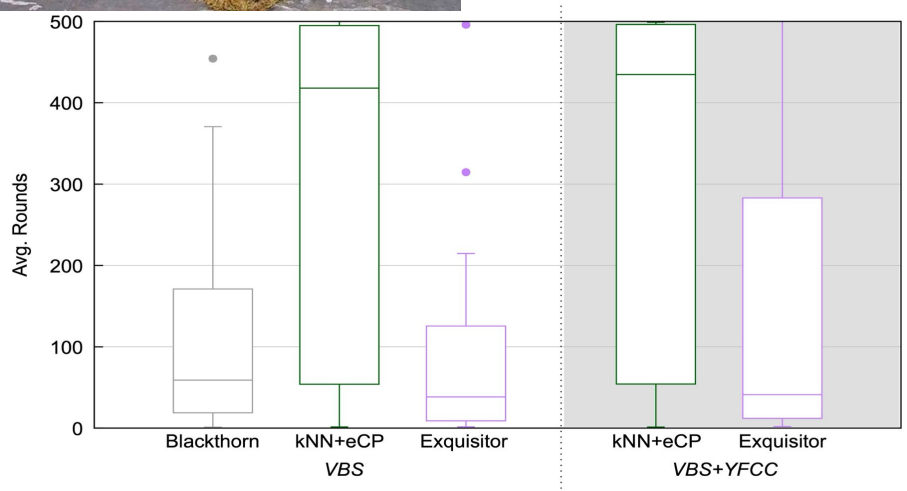


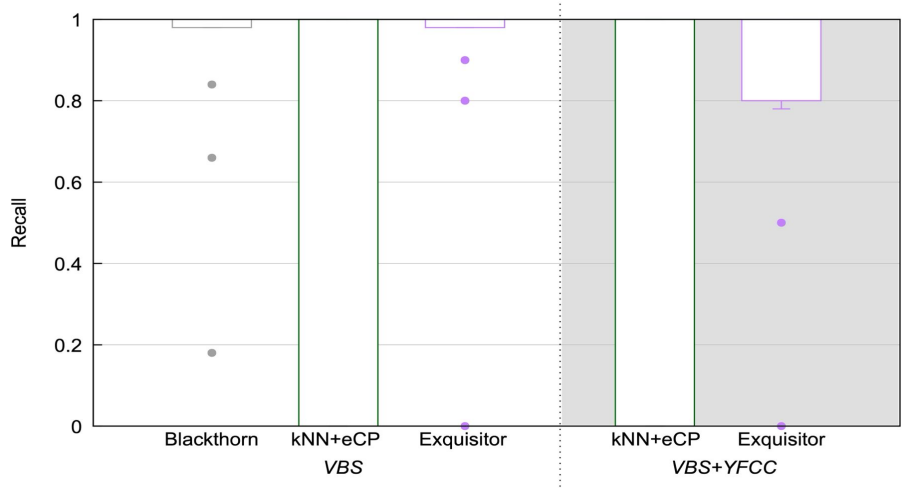


VBS 2020 + YFCC 100M

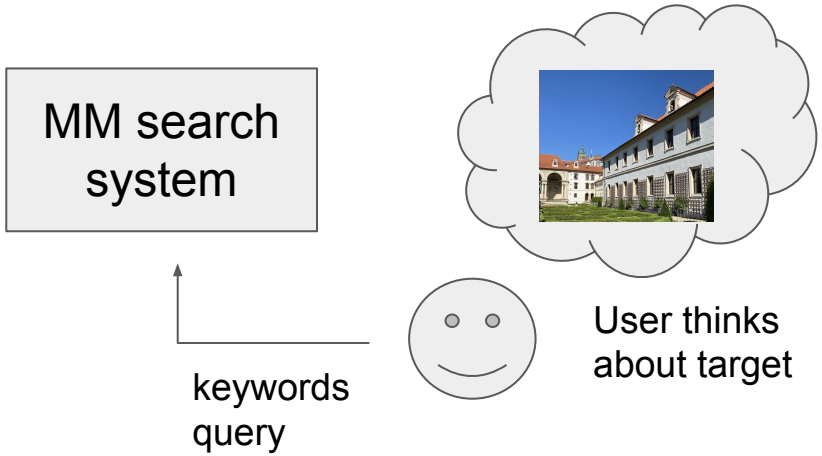
Needle = Solving VBS tasks
Haystack = YFCC 100M (= 100x VBS)

Quality is *nearly* the same



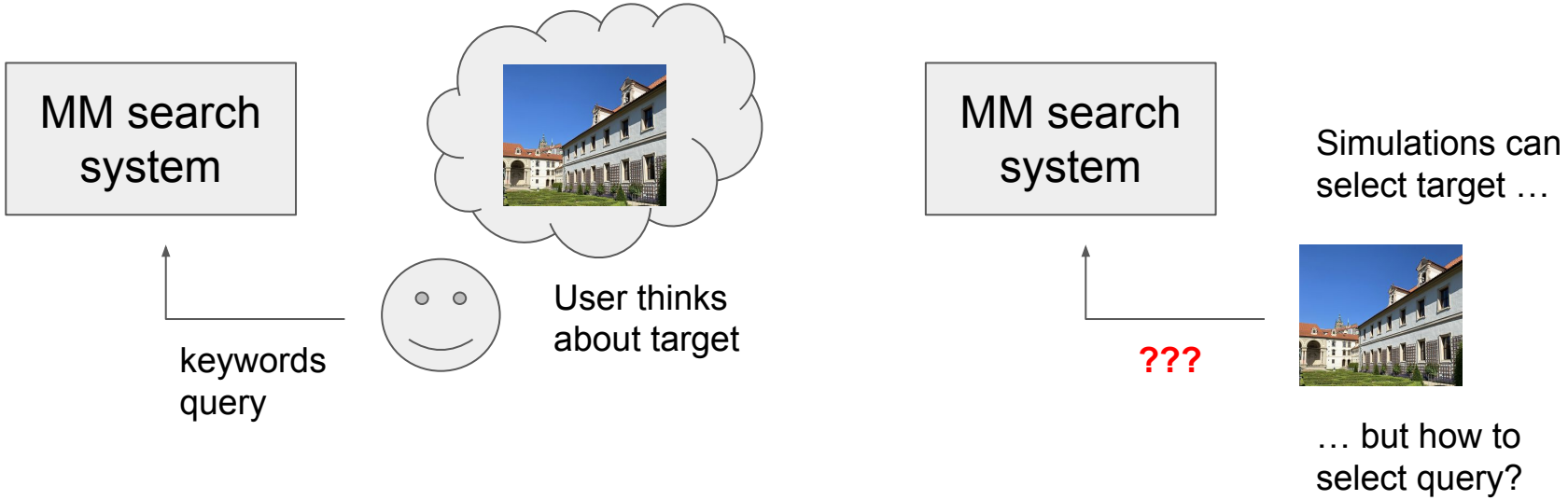


Can we simulate a keyword KIS query?



L. Peška et al: Towards Evaluating and Simulating Keyword Queries for Development of Interactive Known-item Search Systems. ICMR 2020: 281-285

Can we simulate a keyword KIS query?



L. Peška et al: Towards Evaluating and Simulating Keyword Queries for Development of Interactive Known-item Search Systems. ICMR 2020: 281-285

Can we simulate a keyword KIS query?

For simulations of artificial users, a searched target image TI is available.
The query could be extracted from the image!

Tested hypothesis - can algorithm select classes from TI such that the query performance is (on average) similar as for real user queries over TI?

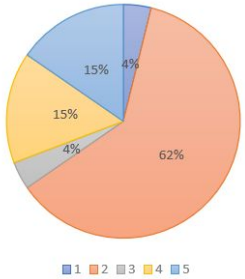
We performed a following study with the following keyword search system

- Each image received deep feature vector from GoogleNet,
  each dimension of the vector = **one class label**
- Query = list of classes = list of dimensions => aggregation of dim values

L. Peška et al: Towards Evaluating and Simulating Keyword Queries for Development of Interactive Known-item Search Systems. ICMR 2020: 281-285

Can we simulate a keyword KIS query?

Toy example to illustrate the query simulation idea:

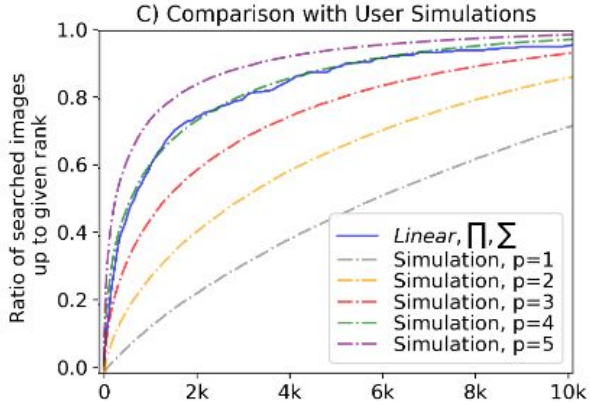


Target image deep vector [0.1, 0.4, 0.1, 0.2, 0.2]

Modified TI with $f(x)=x^p$, p=2 [0.01, 0.16, 0.01, 0.04, 0.04]

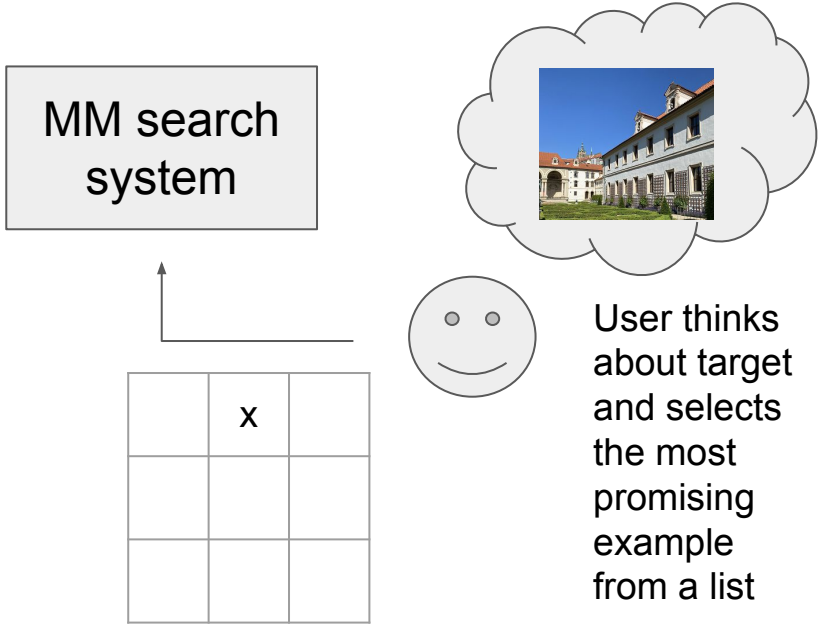
Now the method samples query classes from TI vector based on modified scores

Results compared with real user queries



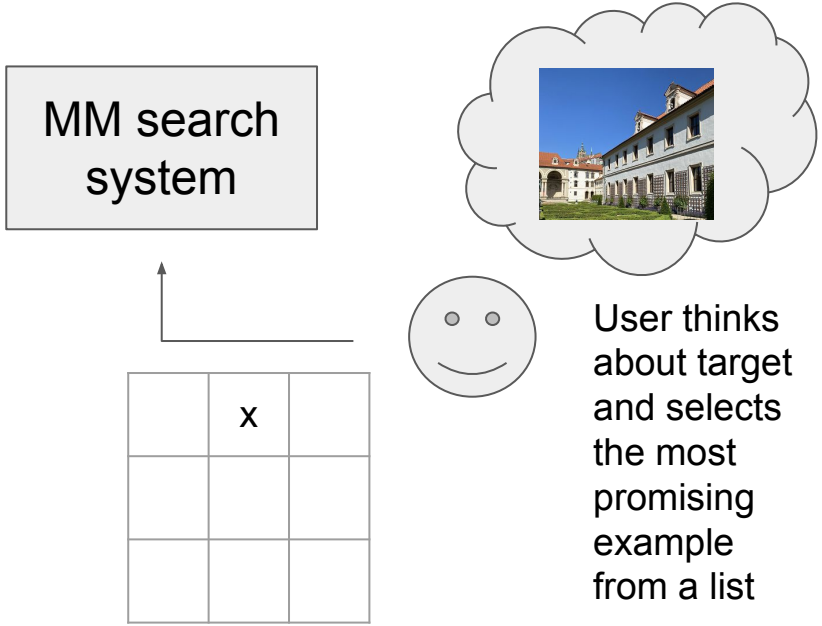
L. Peška et al: Towards Evaluating and Simulating Keyword Queries for Development of Interactive Known-item Search Systems. ICMR 2020: 281-285

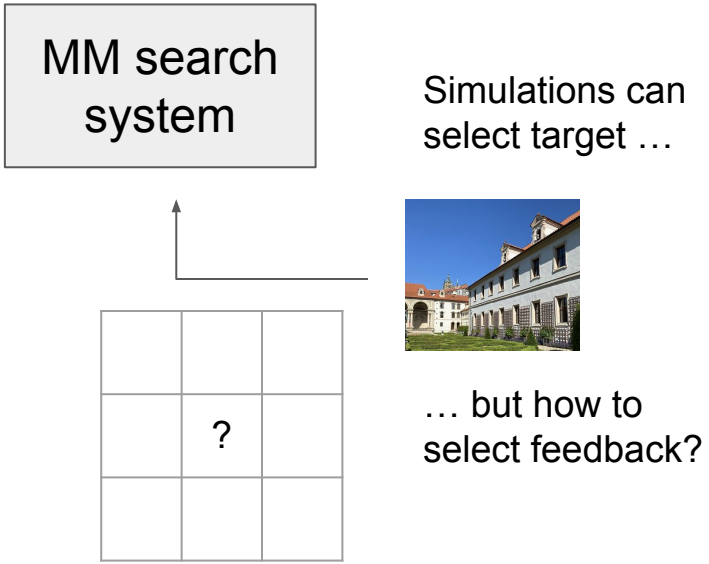
Can we simulate relevance feedback?



L. Peška et al: Evaluating a Bayesian-like Relevance Feedback Model with Text-to-Image Search Initialization, accepted to MTAP, 2022

Can we simulate relevance feedback?





L. Peška et al: Evaluating a Bayesian-like Relevance Feedback Model with Text-to-Image Search Initialization, accepted to MTAP, 2022

Can we simulate relevance feedback?

For simulations of artificial users, a searched target image TI is available.
We can use distances between TI and images in the display!

Tested hypothesis - can algorithm select an example image from the list such that the feedback search performance is (on average) similar as for real user?

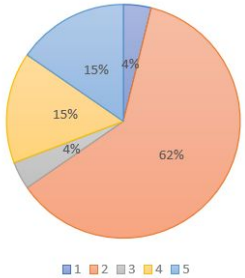
We performed a following study with the following keyword search system

- Each image received deep feature vector from the W2VV++ model
- User selection = selection of a promising image from a list of images

L. Peška et al: Evaluating a Bayesian-like Relevance Feedback Model with Text-to-Image Search Initialization, accepted to MTAP, 2022

Can we simulate relevance feedback?

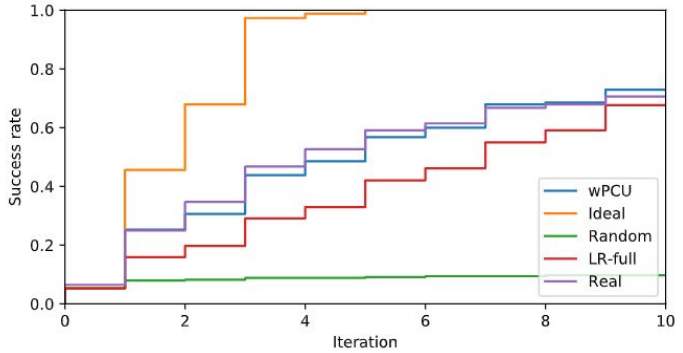
Toy example to illustrate the selection of an image from display D:



Distances from TI to O_i in D [0.1, 0.4, 0.1, 0.2, 0.2]

Modified dist with $f(x)=x^p$, p=2 [0.01, 0.16, 0.01, 0.04, 0.04]

Now the basic idea is to sample images from display based on modified distance scores



Results compared with real user selections

Some other tricks tested, paper has 40 pages!

L. Peška et al: Evaluating a Bayesian-like Relevance Feedback Model with Text-to-Image Search Initialization, accepted to MTAP, 2022

p=12

Summary

Artificial **interactive** users are **cheap**, **fast** and **reproducible**
⇒ they can be useful during **system development** and **parameter tuning**

Artificial users are **not** the same as **real users**
⇒ they can only indicate **likely tradeoffs**

Caveats:

- Artificial users do not have user experiences!

- Artificial users may not properly exercise all aspects of a system!

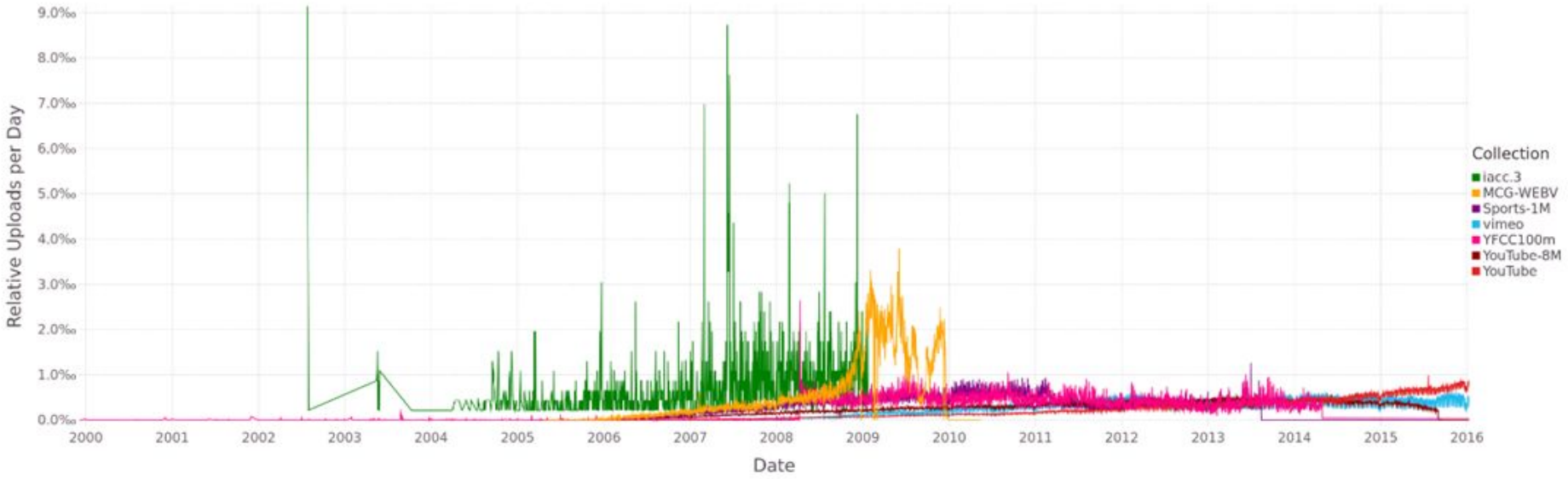
- Simulating real users is very hard!

Real-World Datasets

- **Research needs reproducible results**
  - standardized and free datasets are necessary
  - need to be easily redistributable
  - need to be representative of the ‘real world’
- **One problem with many datasets:**
  - current state of web video in the wild is not or no longer represented accurately by many of them
- **Hence, we also need datasets that model the real world**
  - One such early effort: V3C

Videos in the Wild

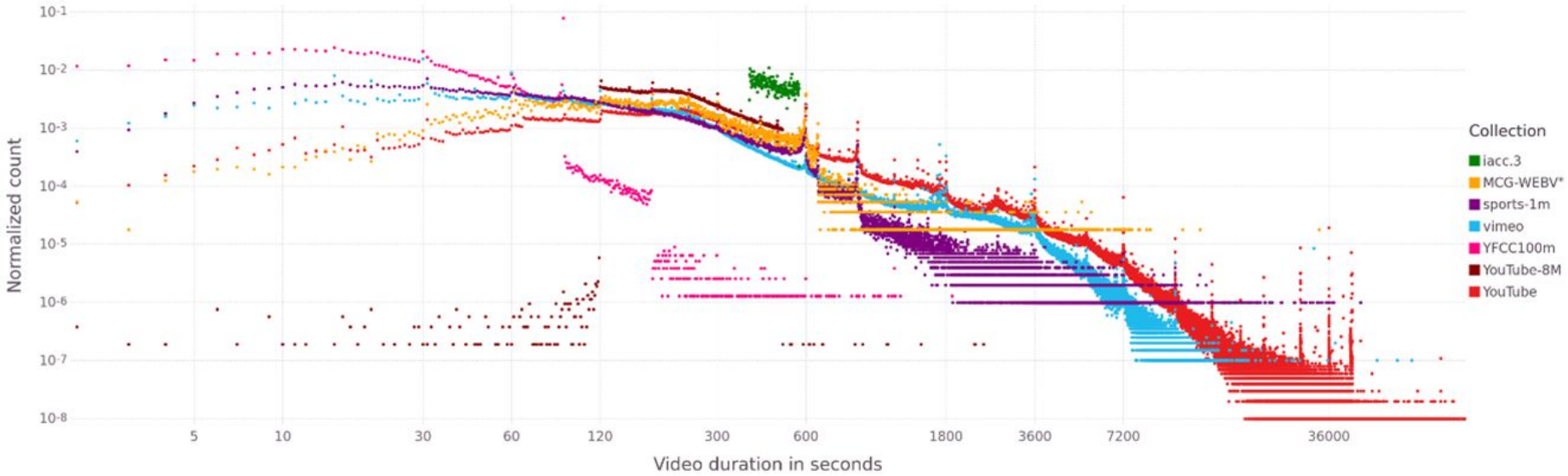
Age-distribution of common video collections vs what is found in the wild



Rossetto, L., & Schuldt, H. (2017). Web video in numbers-an analysis of web-video metadata. arXiv preprint arXiv:1707.01340.

Videos in the Wild

Duration-distribution of common video collections vs what is found in the wild



Rossetto, L., & Schuldt, H. (2017). Web video in numbers-an analysis of web-video metadata. arXiv preprint arXiv:1707.01340.

Available Datasets

- **Past TRECVID data**
  - https://www-nlpir.nist.gov/projects/trecvid/past.data.table.html
  - Different types of usage conditions and license agreements
  - Ground truth, annotations and partly extracted features are available
- **Past MediaEval data**
  - http://www.multimediaeval.org/datasets/index.html
  - Mostly directly downloadable, annotations and sometimes features available
- **Some freely available data sets**
  - TRECVID IACC.1-3
  - TRECVID V3C1 (starting 2019), also be used for VBS (download available)
  - BLIP 10,000 http://skuld.cs.umass.edu/traces/mmsys/2013/blip/Blip10000.html
  - YFCC100M https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67
  - Stanford I2V http://purl.stanford.edu/zx935qw7203

Available Datasets

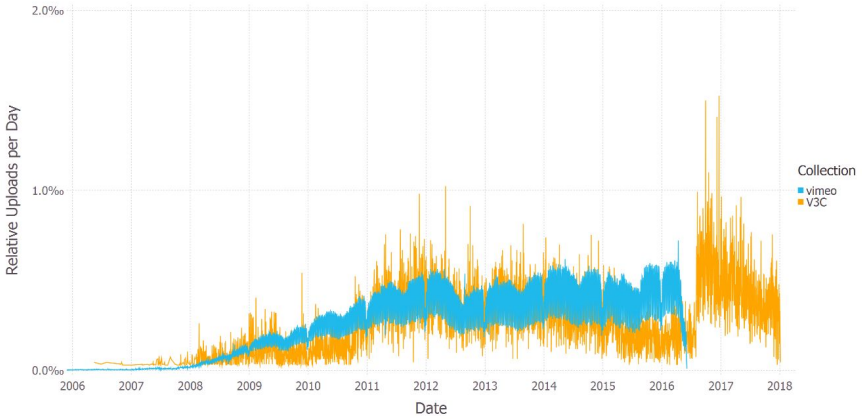
- **MPEG CDVA data set**
  - Mixed licenses, partly CC, partly specific conditions of content owners
- **NTCIR-Lifelog datasets**
  - NTCIR-12 Lifelog - 90 days of mostly visual and activity data from 3 lifeloggers (100K+ images)
    - ImageCLEF 2017 dataset a subset of NTCIR-12
  - NTCIR-13 Lifelog - 90 days of richer media data from 2 lifeloggers (95K images)
    - LSC 2018 - 30 days of visual, activity, health, information & biometric data from one lifelogger
    - ImageCLEF 2018 dataset a subset of NTCIR-13
  - NTCIR-14 - 45 days of visual, biometric, health, activity data from two lifeloggers

Vimeo Creative Commons Collection

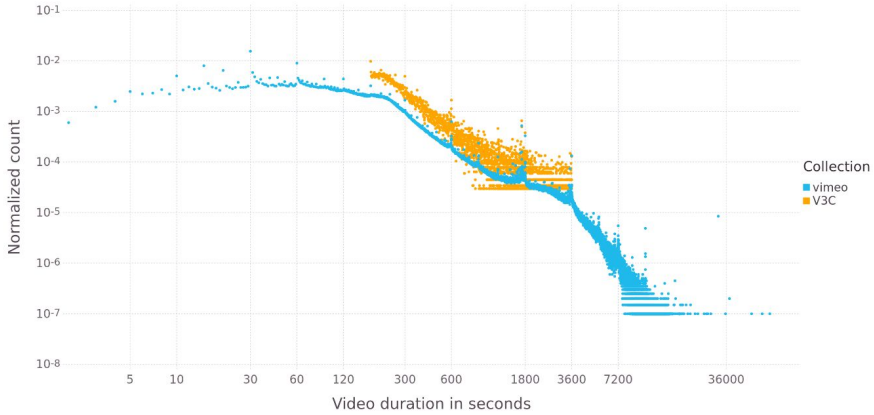
The Vimeo Creative Commons Collection (V3C) consists of ‘free’ video material sourced from the web video platform vimeo.com. It is designed to contain a wide range of content which is representative of what is found on the platform in general. All videos in the collection have been released by their creators under a Creative Commons License which allows for unrestricted redistribution.

Partition	V3C1	V3C2	V3C3	Total
File Size	2.4TB	3.0TB	3.3TB	8.7TB
Number of Videos	7’475	9’760	11’215	28’450
Combined Video Duration	1000 hours, 23 minutes, 50 seconds	1300 hours, 52 minutes, 48 seconds	1500 hours, 8 minutes, 57 seconds	3801 hours, 25 minutes, 35 seconds
Mean Video Duration	8 minutes, 2 seconds	7 minutes, 59 seconds	8 minutes, 1 seconds	8 minutes, 1 seconds
Number of Segments	1’082’659	1’425’454	1’635’580	4’143’693

V3C Uploads and Duration



Age-distribution of the V3C in comparison with vimeo data



Duration-distribution of the V3C in comparison with vimeo data

V3C Overview

V3C Content

- Original Videos
- Video metadata from vimeo
- Automatically generated video shot boundaries
- Lossless video keyframes for every segment
- Thumbnail image for every keyframe

#00001

#00072

#00314

#00885

#01411

#01976

Challenges with V3C and beyond

- **Distribution of content of videos on the Web changes over time**
  - → new datasets are required every few years to stay representative
- **Copyright limits content diversity**
  - → little to no creative-commons movies, movie trailers, music videos, etc.
- **Size matters**
  - → too small datasets don't represent realistic content diversity
  - → too large datasets become infeasible to handle
- **One size does not fit all**
  - → despite being generally representative, some different content distributions might be needed for different tasks

Evaluation Setting

- **Video Browser Showdown setting**
  - same place
  - same time
  - same conditions
  - same tasks
  - same metrics

- **Limitations due to travel restrictions**
  - same tasks, same metrics, same time
  - similar conditions, different places
  - how can this be overcome?

Distributed Retrieval Evaluation Server (DRES)

- Designed to support distributed evaluation of interactive multimedia retrieval
- Interaction through Browser and OpenAPI
- Supports various media types
- Supports diverse task settings
- Provides real-time evaluation feedback
- Designed to be extendable to future use-cases
- Open-source, available via https://dres.dev

Distributed Retrieval Evaluation Server (DRES)

Measurements and Insights

- **Time to correct submission / ratio of correct vs incorrect submissions as primary measure of retrieval effectiveness**
  - results produced by human/machine collaboration
  - how to disentangle the contributions of human operator performance from machine capabilities?
  - multiple users per system, both experts and novices, provide insight to some degree
  - more insight to be gained by sampling results at intermediate steps
- **Logging the results returned to the user**
  - for every submission, also send the top-n retrieved results
  - dedicated API endpoint, process works without human intervention
  - provides insights about effectiveness of retrieval model vs result presentation / browsing
- **How did those results come to be?**
  - results dependent on query input and machine capabilities
  - due to diversity in querying and interaction options, unified recording is still an open issue

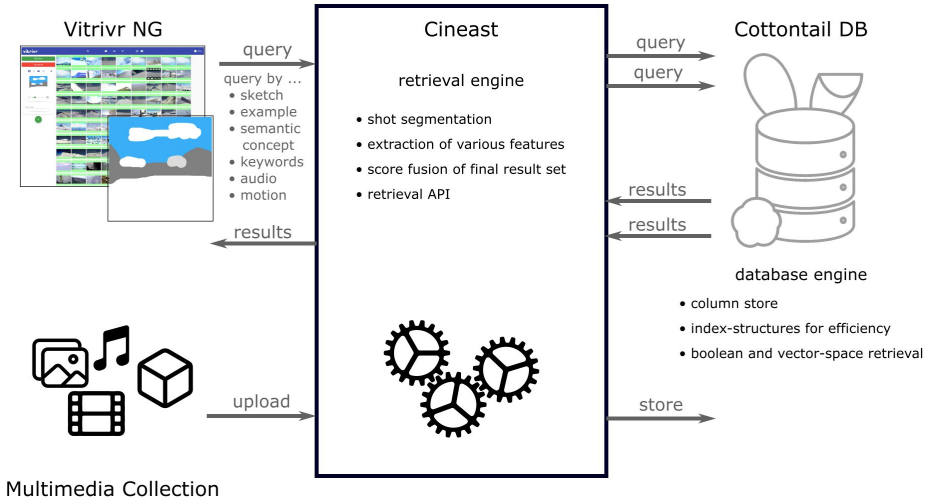
Test Session

Now it’s your turn

- **Several retrieval systems from recent years of VBS**
  - vitrivr
  - CVHunter
  - Exquisitor
  - diveXplore
- **Instance of evaluation server DRES**
- **Tasks akin to VBS**

vitrivr

- **multimedia retrieval stack with support for various media types and query modes**
- **three primary system components**
  - Database layer (Cottontail DB)
  - Querying engine (Cineast)
  - User interface (vitrivr-ng)
- **fully open-source**
  available via https://vitrivr.org



CVHunter

Simple research project combining an interactive interface (WPF .NET) with a deep neural network based on joint embedding

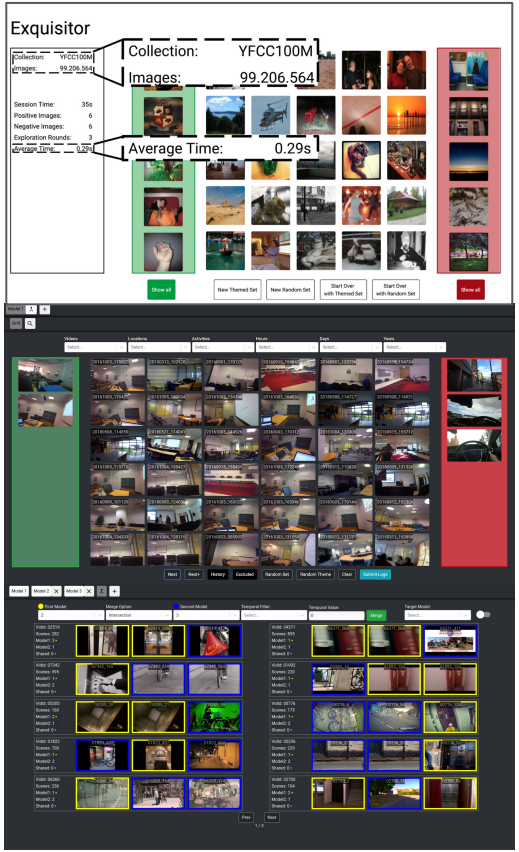
Supports not only basic free form text queries

- Temporal queries, context-aware queries
- Bayesian relevance feedback and its temporal variant
- Various standard browsing options (video summary, temporal context)
- Implements result set logging for performance analysis

Exquisitor

- Large scale interactive learning
  - High-dimensional index (eCP)
  - Query Optimisation Policies
  - Modalities (ImageNet 13K, Kinetics-700)
  - Relevance Feedback (Linear SVM)

- User Interface
  - Web Reactjs
  - Mobile (Kotlin)
  - Cross (React Native)



diveXplore

- **Video units**
  - uniformly sampled shots
  - video summaries (maps)
- **Concepts, classes, events**
  - Concepts: ImageNet, Places365
  - Objects: MS COCO (YOLOv5)
  - Events/Actions
  - Similarity features (CNN hashes)
- **Indexing and Middleware**
  - Data server
  - MongoDB (analysis results)
  - NodeJS query server
- **Web Interface (Angular)**
  - Shot search
  - Map search