The TV Workstation project: a research scope

Keynote talk at the LIFAT seminar

Mathieu Delalandre

University of Tours (UT), LIFAT Laboratory, RFAI group
Tours city, France
mathieu.delalandre@univ-tours.fr
Talk available at http://mathieu.delalandre.free.fr/talks.html

Tours city (France)
10^{th} of July, 2025

Summary

Introduction

TV video capture

Real-time TV video processing

Conclusions and perspectives

Introduction

- Television (TV) is a huge source of multimedia data[1],
  - $\simeq 27,000$ channels worldwide,
  - $\simeq 55\%$ in Europe, Russia, China, USA,
  - provided with DTT, SaT, Cable TV, IPTV and InternetTV,
  - e.g. France / Vietnam ($\simeq 210$ channels), USA ($\simeq 1,760$ channels),
- Computer Vision and AI could be applied to TV,
  - Social TV, Sync2Ad, Fact-Checking, GenAI for TV, ...,
- A Workstation has to support the scalability / real-time issues, this leads us to develop the TV Workstation since 2017.

[1] audio/video & metadata

Summary

Introduction

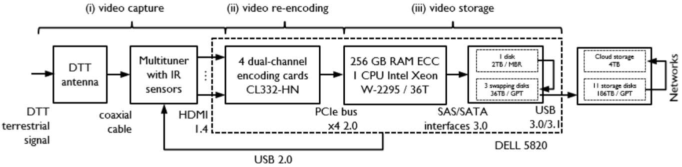
TV video capture

Real-time TV video processing

Conclusions and perspectives

The DELL 5820 computer and tool suite (1/2)

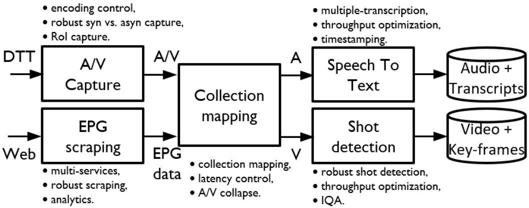
The DELL 5820 **computer** processes 8 channels (HD, 30 FPS, 24h/day), with real-time audio / video (A/V) encoding, control of tuners with IR sensors, internal / external storage of $38 + 190$ TB.



Resolution		Audio/ Video	CPU rate	Video Mbps	TB/ month	Audio Kbps	GB/ month
HD	1280 × 720		20 %	3	7.23	256	621
SD	720 × 576	asyn	12 %	1.6	3.89	160	384
Low	320 × 240		8 %	0.56	1.36	128	308

The DELL 5820 computer and tool suite (2/2)

The DELL 5820 **computer** is offered with a tool suite for adaptive capture, mapping and first analysis of A/V data.



Sources	Area	Length	Size	Ch	BCE	Col	Desc	Words
3	Francophonie	$\simeq$ 2 years	160 GB	310	5 M	120.2 k	1 M	69 M

Ch, **BCE**, **Col**, **Desc** stand for channels, broadcast events, collections and descriptions.

Whisper model	tiny	base	small	medium	large
Throughput	39.8	30.1	6.5	$\simeq$ 4.5	$\simeq$ 2
Audio (h) / week	6,685	5,055	1,090	$\simeq$ 755	$\simeq$ 335
Memory (GB)	22.5	27.5	48	$\simeq$ 115	$\simeq$ 192
CPU rate	$\simeq$ 90 %				

Partial video copy detection (1/4)

Partial video copy detection (PVCD) aims at finding short segment(s) which have transformed in long video(s):

- it is a key topic with application domains (copyright, retrieval),
- existing datasets (VCDB, FIVR-PVCD, VCSL) offer no scalability, control of spatial degradations, null latency and frame-level annotation,
- a TV-based protocol was proposed to design the STVD-PVCD dataset on the task, public available[2,3] [**ORASIS2021**,**ICIAP2022**].

[2] `https://dataset-stvd.univ-tours.fr/pvcd/`

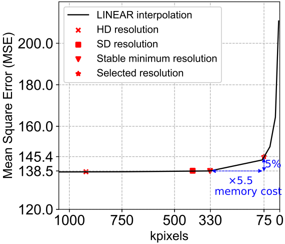
[3] e.g. cove.thecvf.com, datasets.visionbib.com, homepages.inf.ed.ac.uk, kaggle.com, opendatalab.com, paperswithcode.com, . . .

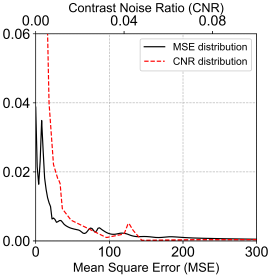
Partial video copy detection (2/4)

STVD-PVCD is the best dataset for scalability and noise control.
It uses a strategy for memory cost optimization[4].

Datasets	**VCDB** 2016	**FIVR-PVCD** 2021	**STVD** 2021	**VCSL** 2022
References	28	100	**243**	122
Positive videos	528	N/A	**19,280**	9,207
Positive pairs	9k	10,8k	**1,688k**	281k
Negative videos	**100,000**	N/A	64,040	N/A
Duration (h)	2,030	N/A	10,660	**17,416**
Noise characterization	real noise	real noise	**noise-free**	real noise
Timestamping (s)	1	1	$\frac{1}{30}$	1

(h): hours, (s): seconds, N/A: not available, k: thousands





[4]75k pixels = 320 × 240 at 560 kbps

Partial video copy detection (3/4)

STVD-PVCD allows a fine characterization for PVCD:

- a root capture plus 5 test sets,
- *video cut* is controlled with a latency model,
- *downscaling*, *compression* depend of correlated parameters,
- *flipping*, *rotating*, *black-border insertion* are standard,
- *video speeding* is done with **[ICAIIC2020]**.

	Video cut	Downscaling	Compression	Flipping	Rotating	Black-border	Video speeding
Root capture	•						
Hello World	•	•	•				
Pixel attacks	•	•	•				
Global transforms	•	•	•	•	•	•	
Video speeding	•	•	•				•
Combination	•	•	•	•	•	•	•

Root capture

Hello world

Pixel attack


Global transforms


Video speeding


Combination

Partial video copy detection (4/4)

STVD-PVCD is suitable to characterize PVCD methods.

- Detection: key-frame based method[5] **[ICPR2016]**
- Features: 10 (BRIEF, 9 CNN features)
- Dataset: STVD sampling[6] without/with training[7]
- Experiments: $>$ 4.4 M vectors and $>$ 445 B matchings
- Metric: F_1

	BRIEF	VGG-16
Hello world	**0.98**	N/A
Pixel attack	0.59	**0.64**

Hello world & Pixel attacks - BRIEF / CNN feature - without training

	Last FC	MAC	R-MAC
ResNet50-v1	**0.926**	0.828	0.823
Inception-v1	0.923	0.738	0.782
VGG-16	0.894	0.922	0.918

Gobal transforms - 9 CNN features - with training

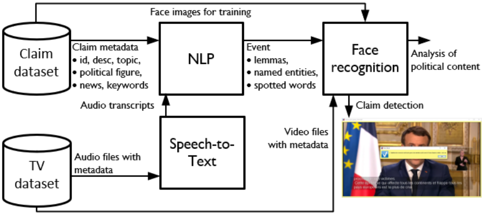
[5] Cosine similarity, at least one key-frame detected

[6] large-scale, balanced positive/negative distribution, accurate timestamping

[7] ratios of $\frac{3}{5}$ for training and $\frac{2}{5}$ for testing

Multimodal audio/video analysis for fact-checking

Fact-checking checks the veracity of claims from various media (print, TV, radio, Web, SN). There is none A/V dataset, we have designed the large-scale french STVD-FC:



- containing 6,730 news / political TV programs (6,540 h) of the French presidential election 2022[8] ($\simeq$ 50 Mwords, $\simeq$ 706 Mimages, 1.96 TB),
- linked to $\simeq$ 10,000 claims ($\simeq$ 10 years / height web services) [**ISS2025**],
- public available[9] [**CBMI2022**] tested a first system [**VISAPP2024**].

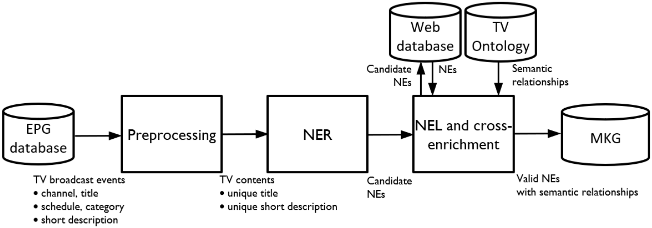
[8] 1^{st} of February to 1^{st} of May 2022

[9] `https://dataset-stvd.univ-tours.fr/fc/`

Multimedia Knowledge Graph (MKG) (1/2)

Multimedia Knowledge Graph (MKG) represents multimedia content (text, image, A/V, etc). There is no large-scale French/MKG, we have designed STVD-KG:

- a preprocessing maps TV broadcast events into contents,
- NEs are detected with NER Spacy/casEN intersection for robustness,
- NEL processes candidate named entities with WebDB and TV ontology,
- STVD-KG is $\times 10$ bigger than state-of-the-art, public available[10].



NER, **NEs**, **NEL**, stand for Named Entity Recognition, Named Entities and Named Entity Linking.

EPG	BCE	CoI	NEs	Triples	Properties
1 year	2.9M	70k	84k / 5.7M	27.6M	21

EPG, **BCE**, **CoI**, **NEs** stand for Electronic Program Guides, broadcast events, collection and named entities.

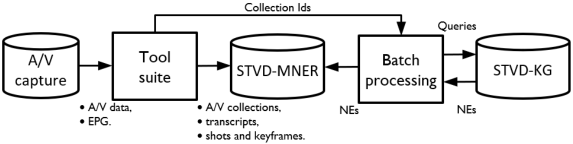
[10] https://zenodo.org/records/15368241

Multimedia Knowledge Graph (MKG) (2/2)

MKG is inherently a Multimodal Knowledge Graph (*M*KG). *M*KG construction is related to the Multimodal Named Entity Recognition (MNER) in A/V data with semi-supervised learning.

LOC: external vs. internal PEO: old vs. young ORG: building vs. object

MNER for A/V is specific, a STVD-MNER$_\beta$ dataset is proposed.



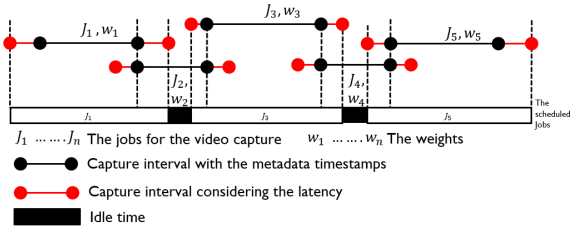
Dur	Col	A/V files	Transcripts	Audio	Video	NEs
819h	284	843 $\times$2	yes	256 kbps	0.56Mbps 320 $\times$ 240	1,231

Dur, **Col**, **NEs** stand for duration, collection and named entities.

Parallel machine scheduling (PMS) for A/V capture

Problem statement: large-scale capture has an hardware / memory cost not needed[11]. A partial capture with PMS:

- is an off-line / no preemptive scheduling using static execution times,
- is a Weighted Interval Selection Problem (WISP) NP-hard having polynomial approximation algorithms (e.g. $GREEDY_\alpha$ [**JA2003**]),
- has a latency $L(t)$ as key parameter of the scheduling problem,
- is delivered with public available dataset STVD-PMS[12] (170 days, 26 channels, 99k jobs, 5,615 hashcodes, offline/online latency).



[11] Frequent, political content, rich EPG data, …

[12] https://dataset-stvd.univ-tours.fr/pms/

Summary

Introduction

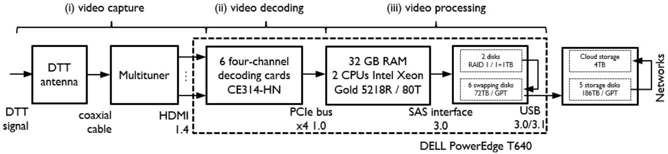
TV video capture

Real-time TV video processing

Conclusions and perspectives

The DELL PowerEdge T640 computer

The DELL PowerEdge T640 **computer** processes 24 channels for real-time video decoding and processing with high-performance CPUs[13] and having an internal / external storage of $72 + 190$ TB.



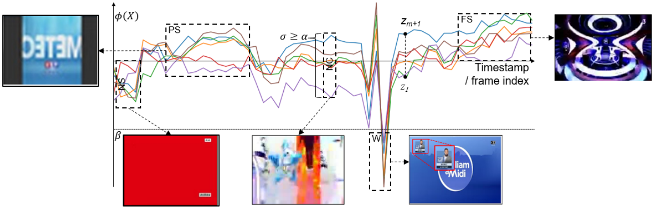
Ch	BPP	Res	FPS	Images	Bandwidth		
24	32	SD	$600 = 24 \times 25$	51.8 M/day	0.69 GB/s	57.9 TB/day	34%
		HD	$528 = 24 \times 22$	45.6 M/day	1.81 GB/s	152.9 TB/day	91%
		Full HD	$240 = 24 \times 10$	20.7 M/day	1.85 GB/s	156.4 TB/day	93%

[13] 2×40 threads with AVX 512 Vector Neural Network Instructions

Real-time PVCD

Real-time PVCD processes with a deadline Δ (e.g., 1-3s) and can be applied to multiple video streams [**CBMI2021**,**CAIP2023**]:

- with real-time video decoding using hardware on the Workstation,
- with rigid (ZNCC) and no-rigid (2D CNN) features for matching,
- with key-frame selection methods using goodness criteria.



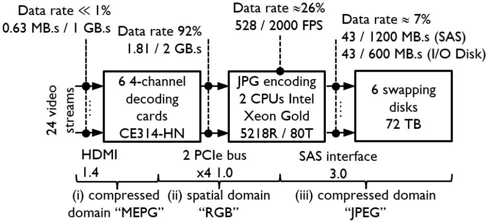
Time optimization for real-time deep learning to investigate:

- acceleration [14] with INT8 and VNNI [**CCIS2020**],
- soft real-time with adaptive inference [**PR2020**].

[14] $\simeq \times 15$ acceleration on *ResNet-50* (OpenVino vs. TensorFlow).

Real-time frame capture and IQA (1/2)

Real-time frame capture decodes videos into frames re-encoded as image files (e.g. jpeg). The workstation can process 24 streams (22 FPS / HD) in real-time and offers a large storage (72 TB).



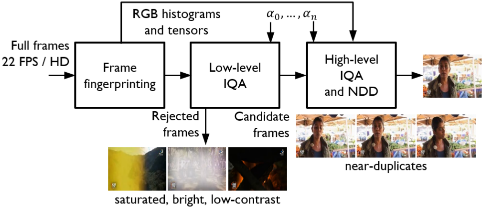
No bottleneck, the problem is the storage cost (3 weeks max).

	Day	Month	Year	Disks
data	3.4 TB	103.2 TB	1.22 PB	72 TB
image	45.6 M	1.4 B	16.7 B	0.96 B

M, B, TB, PB stand for Millions, Billions, Terabyte, Petabyte

Real-time frame capture and IQA (2/2)

Image Quality Assesment (IQA) filters high quality frames into a two steps pipeline.



- low-level IQA filters out low quality frames with standard processing,
- high-level IQA requires time-efficient blur detection methods [**CIS2023**],
- Near-Duplicate Detection (NDD) filters out duplicate frames for storage,
- parameters $\alpha_0, \ldots, \alpha_n$ are set for storage requirements (e.g. $\simeq 12$ FPM).

Summary

Introduction

TV video capture

Real-time TV video processing

Conclusions and perspectives

Conclusions and perspectives

- project launched in 2017, specific / ready-to-use platform,
- cross-disciplinary project (CV, NLP, OR),
- 9 researchers working on, $\simeq$ 44.7 k€ of investment,
- 3 PhD (V.H. Le, H.G. Vu, L. Nguyen),
- 7 publications[15] and 4 public datasets STVD[16],
- research perspectives (*M*KG, MNER, RT CV, ...),
- project submission (Fact-Checking, Social TV, TV GenAI).

15 **[AI4TV2019**, **CBMI2021**, **ORASIS2021**, **ICIAP2022**, **CBMI2022**, **CAIP2023**, **VISAPP2024]**

16 `https://dataset-stvd.univ-tours.fr/`

References I

- **[JA2003]** T. Erlebach and F.C.R. Spieksma. Interval selection: Applications, algorithms, and lower bounds. Journal of Algorithms, vol. 46(1), pp. 27-53, 2003.
- **[ICPR2016]** Zhang, Y. and al. Effective real-scenario video copy detection. International Conference on Pattern Recognition (ICPR), pp. 3951-3956, 2016.
- **[AI4TV2019]** M. Delalandre. A Workstation for Real-Time Processing of Multi-Channel TV. Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV), pp. 53-54, 2019.
- **[CCIS2020]** E.P. Vasiliev and al. Performance Analysis of Deep Learning Inference in Convolutional Neural Networks on Intel Cascade Lake CPUs. Mathematical Modeling and Supercomputer Technologies (MMST), Communications in Computer and Information Science (CCIS), vol. 1413, 2020.
- **[PR2020]** N. Passalis and al. Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits. Pattern Recognition (PR), vol. 105, pp. 107346, 2020.
- **[ICAIIC2020]** D. Lee and al. Prediction of network throughput using arima. International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 1–5, 2020.
- **[ICASSP2020]** Y. Huang and al. Leveraging unpaired text data for training end-to-end speech-to-intent systems. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020
- **[IEEE2021]** M. Rafiq and al. Video Description: Datasets & Evaluation Metrics. IEEE Access, vol. 9, 2021.
- **[CBMI2021]** V.H. Le, M. Delalandre and D. Conte. Real-time detection of partial video copy on TV workstation. Conference on Content-Based Multimedia Indexing (CBMI), pp. 1-4, 2021.
- **[ORASIS2021]** V.H. Le, M. Delalandre and D. Conte. Une large base de données pour la détection de segments de vidéos TV. Journées Francophones des Jeunes Chercheurs en Vision par Ordinateur (ORASIS), 2021.
- **[CBMI2022]** F. Rayar, M. Delalandre and V.H. Le. A large-scale TV video and metadata database for French political content analysis and fact-checking. Conference on Content-Based Multimedia Indexing (CBMI), 2022.

References II

- **[ICIAP2022]** V.H. Le, M. Delalandre and D. Conte. A large-Scale TV Dataset for partial video copy detection. International Conference on Image Analysis and Processing (ICIAP), Lecture Notes in Computer Science (LNCS), vol. 13233, pp. 388-399, 2022.
- **[CAIP2023]** V.H. Le, M. Delalandre and H. Cardot. Performance characterization of 2D CNN features for partial video copy detection. International Conference on Computer Analysis of Images and Patterns (CAIP), Lecture Notes in Computer Science (LNCS), vol. 14184, pp. 205-215, 2023.
- **[CIS2023]** X. Wang and X. Liang and S. Li and J. Zheng. Efficient image blur detection via hierarchical edge guidance and region complementation. Journal: Complex & Intelligent Systems, 2023.
- **[VISAPP2024]** F. Rayar. Fact-checked claim detection in videos using a multimodal approach. Conference on Computer Vision Theory and Applications (VISAPP), 2024.
- **[ISS2025]** F. Rayar, J. Nicey. Lebonfait ? Retour sur la création d'une base de faits vérifiés en français. Journées Infox sur Seine (ISS), 2025.