HAL
open science

A Large-Scale Audio and Video Fingerprints-Generated Database of TV Repeated Contents

Jean-Hugues Chenot, Daigneault Gilles

To cite this version:

Jean-Hugues Chenot, Daigneault Gilles. A Large-Scale Audio and Video Fingerprints-Generated Database of TV Repeated Contents. 12th International Workshop on Content-Based Multimedia Indexing (CBMI2014), Jun 2014, Klagenfurt, Austria. pp.02-02, 10.1109/CBMI.2014.6849818 . hal-01017118

HAL Id: hal-01017118

https://hal.science/hal-01017118

Submitted on 16 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Large-Scale Audio and Video Fingerprints-Generated Database of TV Repeated Contents

Application of lightweight fingerprints to large-scale semi-supervised structuring of TV contents, over 4 years and 10 TV channels

Jean-Hugues Chenot, Gilles Daigneault
Institut National de l'Audiovisuel
Bry sur Marne, France

***Abstract*—Using specifically-designed lightweight audio and video fingerprints, we were able to detect repeated contents over a quasi-uninterrupted recording of 10+ TV channels, over more than 4 years, starting January 2010 (380,000 hours); the detection independently uses audio and video fingerprints. The results are stored into a database that holds more than 20 million detected repeats. Detections range from a few seconds up to one hour. The database can be explored using a standard web browser. There are many potential applications, e.g. for structuring and documenting contents.**

Keywords—TV repeats, audio and video fingerprint, multimedia data mining.

I. INTRODUCTION

Since the year 2000 considerable efforts were dedicated by the research community towards near-copy detection for audio and video contents. Products appeared in the industry in 2002 [1], making it possible to recognise musical excerpts from a database of hundreds of thousands of songs, and in video as from 2007[1]. Near-copy detection uses acoustic (for audio) and/or visual (for video) fingerprints. Fingerprints are a compact representation of the original contents, which can be easily compared to detect copies. Usually a set of *candidate* fingerprints is compared with a *reference fingerprints base*, identifying the matching contents. We use video and audio fingerprints to detect and analyse how TV contents and soundtracks are repeated over 10 channels. Every day new fingerprints are generated from daily recordings, and compared with the database of past fingerprints. The resulting database of repeated contents now covers more than 4 years on 10 TV channels, both in audio and video.

II. PREVIOUS WORK

A. Robustness and Discriminance

Considerable work has been dedicated to defining audio and video fingerprints that would present a good robustness to signal alterations (audio filtering, level, compression, video picture framing, zoom, contrast, gamma, rotation...). In the video domain, Yeh [2] distinguishes between global descriptors, based on colours or shot durations, e.g. [2],[3],[4] that are simple but lack discriminance and robustness, and fingerprints relying on local descriptors, e.g. [5],[6], that improve the results, but are heavier and more difficult to use on a large scale. Law-To [7] compared the performances of several video contents fingerprints, most of them using picture content as a primary feature. Other works have improved the robustness to a number of distortions, e.g. zoom, rotations [8]. Shang exploits ordinals relations [9]. Herley [10] and Covell [11] use both audio and video signals to improve speed and robustness.

B. Search Efficiency

A number of authors describe how to be able to search large *fingerprints bases*. Joly [6] showed that search efficiency could be improved for relatively low dimensionalities (typically 16 to 20), by relying on grouped approximate k-nearest neighbours search. Poullot [12] relies on embedding multiple descriptions of picture contents into relatively low-dimensional descriptors. Higher dimensions are still searchable, e.g. using principal component analysis (PCA), random projections and locally-sensitive hashing [13] or product quantisation hashing [14] to reduce or handle the number of dimensions. Unlike in the audio domain, we have found few published research works [12] stating that fingerprints could be used in the video domain to detect near-copies over datasets beyond 100,000 hours. Beyond fingerprint compactness or search efficiency, this may be due to other factors such as the difficulty to record such amounts of contents.

C. Previous Work at INA

INA started developing fingerprint technologies in the year 2000, and successfully implemented systems that have been in use since 2005, with two main applications: detection of broadcasts of INA-originated contents, and filtering of contents on UGC (User-Generated Contents) sites, to manage rights and revenue sharing on the incoming stream of the sites[1]. The technology described in this paper is different, and it is specifically designed for the application described here.

III. LIGHTWEIGHT AUDIO AND VIDEO FINGERPRINTS

The audio and video fingerprints used in our experiment are improved versions of the video *Temporal* fingerprint mentioned in [7]. The search for lightweight fingerprints was

This work was partially supported by the 2009-2012 FP7-ICT- 231161 PrestoPRIME project.

[1] Dailymotion started using INA's Signature in 2008 : http://www.tvover.net/2008/02/26/Dailymotion+Implements+INA+Technology+For+Detection+Of+Copyrighted+Video.aspx

triggered by our aim to be able to address, by construction, very large-scale applications.

A. Fingerprint Design

The recognition is required to perform well on copies and excerpts of the same video or audio sequences, even when the sequences have gone through different processes and distortions (e.g. editing, level/colour/gain, geometric transformations…). Beyond some level of distortion, detecting a copy becomes very difficult, but such extreme cases are rare in practice, and the objective here is not to achieve 100% recall, but rather to provide a good quality of detections, with little or no false detections, even on large amounts of contents. In addition to the traditional robustness, discriminance, and precision concerns, both audio and video fingerprints have to be easily computed and stored. To facilitate the search, we also added the requirement of a uniform distribution in fingerprints space. This led us to attempt using a global fingerprint exploiting only the temporal activity as the sole source signal (envelope for audio, summed temporal pixel luminance variations for video). Provided that a number of precautions, described below, are taken, this results in surprisingly robust and efficient audio and video fingerprints.

B. Timestamps and Key Frames

For video, the activity is sampled at 25 Hz. The value of a sample is the sum of the squared luminance differences of each pixel with the corresponding pixel in a subsequent picture, weighted by a factor decreasing on the picture borders. This adds up to one activity sample per frame. For audio, the activity is the envelope: the absolute value of the signal, filtered and sub sampled at 100 Hz. There is no need for normalisation, as the fingerprints are phase-based. The fingerprint is only computed around selected *timestamps* that are local maxima of filtered activity over a 5-second temporal window. In video, *timestamps* typically take place at shot boundaries, but also when any action takes place within the picture (gestures, camera motion, flashes, other motion…). In audio, *timestamps* mark phonemes, music notes, percussions, sounds… The local activity maxima are relatively precise and stable markers, as e.g. in Fig. 1.

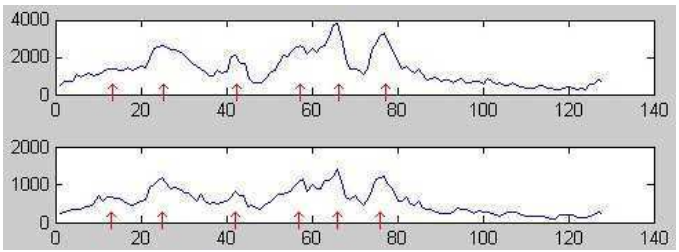

Fig. 1. Video activity and timestamps (red arrows) on two different copies of the same content.

C. Using Phases to Achieve Robustness

Around the chosen *timestamp*, the Discrete Fourier Transform (DFT) of the *log* of the activity is computed over a five-second window, using the Hann window $W(t)$. The small positive constant ε ensures that the *log* is always defined:

$$F(f_i)_{1\leq i\leq n} = DFT_i(\, W(t) * \log10(\, activity(t) + \varepsilon\,)\,) \qquad (1)$$

Using the *log* boosts the smaller activities. The phases of the n first non-null frequencies are retained. The phases capture temporal information relative to a precisely-positioned *timestamp*; they are very robust to signal amplitude distortions introduced by changes in geometry (zoom/stretch/rotation) or dynamics (gain, contrast, gamma, audio compression), that principally affect the modulus. Minor inaccuracies on the position of the *timestamp*, and potential temporal filtering introduced, for example, by changes in frame rate or noise reduction, only have an important effect on the phases of highest frequencies, which are discarded. Phases being defined in a periodic space, using Euclidean distance in [0, 2π] phase space would break continuity; we therefore use a folding function that preserves both continuity and density, mapping continuously and evenly the phase space to [0,255], and quantising each phase into one 8-bit word:

$$Q(f_i) = floor(256*(1+atan2(\, Im(F(f_i)),\, |Re(F(f_i))|)/\pi)) \qquad (2)$$

The resulting fingerprints are vectors in an n-dimensional space, n=16. We have verified that fingerprints are quite evenly distributed within the $[0... 255]^n$ hypercube. Each fingerprint consists of 16 bytes plus 12 ancillary data bytes (a reference to the specific file and *timestamp* where the fingerprint was taken).

D. Video vs. Audio Fingerprint Specificities

In video, activity is sampled at 25 Hz. We keep on average 0.8 fingerprints per second. A video fingerprints database occupies ~22 bytes/second, or 80kB/hour. In audio, the envelope is sampled at 100 Hz. We keep on average 1.4 fingerprints per second. An audio fingerprints database occupies ~37 bytes/second, or 135kB/hour.

IV. Comparing Fingerprints

A. Detecting Fingerprints : the Search

Identifying, within a set of *candidate* videos, quasi-copies of contents within a *reference* set, involves comparing the *candidate fingerprints* to the *reference fingerprints*. To do this we collect as much as possible of *candidate fingerprints* (typically up to 100 hours or 300,000 fingerprints), and search in one pass the *reference fingerprints database* for distance-bound k-nearest-neighbours. We obtain, for each candidate fingerprint, a set of up to $k=10,000$ neighbours.

B. Achieving Robustness : the Vote Step

At this stage, only a few of the unordered neighbours found belong to actual repeats. Robustness is achieved within the *vote* step, which sorts results and searches for sequences of retrieved *candidate/reference* fingerprints pairs that:

a) Come from the same candidate/reference pair of files

b) Share the same candidate/reference timecode offset.

c) Are dissimilar (this prevents detecting simple regular patterns)

d) Are frequent enough (no large temporal gap subsists)

e) Are numerous enough (minimum 4)

When such a set of consecutive fingerprints pairs is found, it is stored as a result line; each line mentions the *candidate* and *reference* files, the starting point in each file for the common sequence, and the duration. By tuning the parameters

on the criteria above, we were able to improve the precision to a point where we don't come across false detections any more.

C. Application to the Detection of Repeats < 10,000 hours

Applying our fingerprint search to detecting repeated contents - without prior knowledge of where such repetitions appear - involves assembling a fingerprint database for the whole set, searching the database, using it both as a *candidate* and as a *reference*, and removing the trivial results (same file, null timecode offset). This can be done easily for database sizes under 10,000 hours.

V. Application to Large TV Databases

A. Avoiding Quadratic Explosion

In our case, even using lightweight fingerprints, given the challenge (>10 channels, several years), it would have been impractical to search the fingerprints database directly, without first dividing the candidates in subsets, and taking into account the risk of quadratic explosion: indeed, we have found several cases where similar contents were repeated thousands of times. Trying to search the whole fingerprints database for such repetitions would generate millions of result lines for each case. To avoid the hassle, and to limit the size and the resources necessary for the search, we proceed incrementally, day after day, and limit the fingerprints database increase by retaining fingerprints only from the *fresh contents*: those contents that are not detected as repeated earlier contents.

B. Application to the Detection of Repeats in TV contents

Every day since 1/1/2010, for 10 TV channels, 24 1-hour half-resolution files[2] are recorded. Audio and video fingerprint files are generated for each media file. A 240-hour *daily fingerprints database* is generated. It is then searched against itself. The *daily fingerprints database* is then trimmed, keeping only first occurrences of contents. The *trimmed daily database* is then compared to the *big base*: the fingerprints database of all past days. The *daily fingerprints database* is then trimmed again, and the fingerprints representing only *fresh contents* are added to the *big base*. The *big base* is then z-curve indexed, to be ready for the next day. This process has been operating for more than four years[3] since January 2010.

C. Raw Results Processing

All detected repeats in video and audio are kept as text results files. Before storing them into a database, a simplification is necessary, as searching the *daily fingerprints database* for repeats returns redundant results: when one segment in a candidate file matches a segment in a reference file, a very similar detection is usually also returned, with *candidate* and *reference* roles switched. More generally, segments that are repeated *n* times on the same day tend to produce up to *n*(n-1)* lines of results. Merging such sets into a single *group of repeats* allows us to reduce the daily figures by more than 80% (from 100 million lines down to 19 million over 4 years). When searching the *twice trimmed daily fingerprints base* against the *big base*, results (5.2 million lines) are much less redundant, thanks to the choice made of including in the *big base* only *fresh contents*' fingerprints.

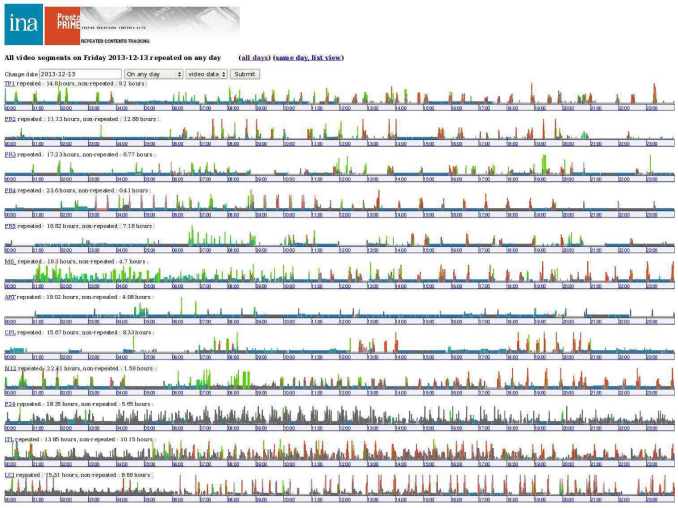

Fig. 2. Daily multi-channel graphical view of repeated contents; time axis is horizontal, 0:00 to 24:00; height of bars represents the number of repeats; coulour reflects the distribution of repeats over days or channels: green if more than one day, red if more than one channel on same day, grey if one channel, one day. The absence of commercials on Arte (7th timeline) and after 20:00 on the French public channels is clearly visible (2nd to 5th timeline).

D. Database Structure

To make the results accessible, they are processed and stored into a mySQL database. To limit the database size and search cost, repeats are represented as *segments* and *groups of repeats*. A *segment* is a repetition of another one when they belong to the same *group of repeats*. This is in line with the choice to include only *fresh contents*' fingerprints into the *big base*. A number of other simplifications are made, e.g. discarding detections that are redundant with existing information, or merging *segments* starting at the same timecode with similar duration, and merging their *groups of repeats* as well. The knowledge resulting from the whole process is stored within two mySQL tables, one for *segments* (34 million), and one for *groups of repeats* (10 million), linked through a *n-to-1* relation.

E. User Interfaces

Searching through such an amount of data is not immediate; we have therefore prepared a number of views, using the phpMyEdit toolbox, to be able to access the mySQL database at different levels. The Yearly view, not shown to save space, presents aggregate figures and allows the user to navigate to a specific day. Figures 2-3 present other views designed to help the user grasp the contents of the database, from the coarsest to the finest level of granularity. On finer views, selectors are available to limit the search (audio/video/both, same/different/any day, same/any channel). When applicable, timelines are provided, showing e.g. in Fig. 2 the number of repeats (height), number of days or number of channels (colour). Clicking on any area on a timeline reduces the search to the specific section. Information collected from the *electronic programme guides* (EPG) is also available on the local timelines views. Thumbnail pictures for the selection give a visual cue, and link to a highly compressed video of the repeated section.

[2] Typically 360x288 or 448x256, 0.5 Mbps with soundtrack.
[3] 4 years in video, starting 1/1/2010; 3.5 years in audio. An 11th channel was added after 11 months, and a 12th in Sept. 2013.

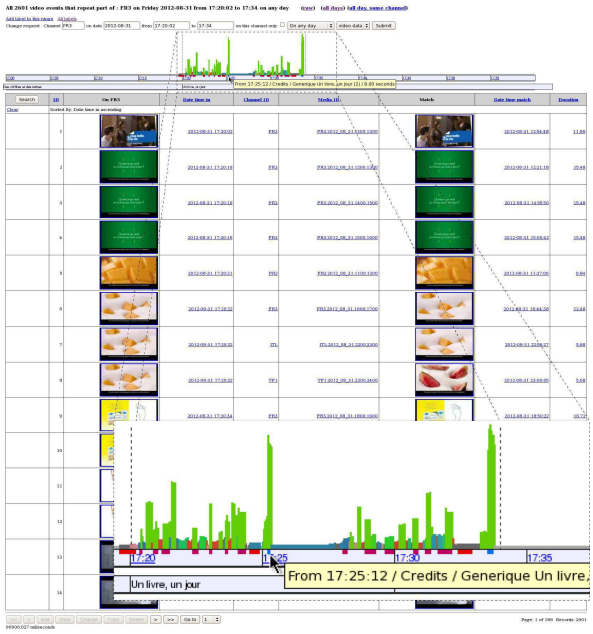

Fig. 3. A view on a 10-minute inter-programme: in that order: a teaser (1st row); a set of commercials; a short 2-minutes litterary programme (not in table, with its title sequence, starting 5 minutes late with respect to the electronic programme guide); a second set of commercials; a teaser; next programme title sequence. On the close-up are marked the already labelled sections, bright red for teasers, red for commercials, blue for credits/titles, cf. Table I.

TABLE I. MANUALLY ENTERED LABEL TYPES AND NUMBERS

Type		*Describing*	*Numbers*
Pub		Commercial	654
Jingle		Very short musical and visual punctuation	77
Credits		Credits & Title sequence	188
Series		Series episode & cartoon	15
Long		Other long programmes	23
Teaser		Announcement for programme(s) to come	122
Excerpt		Excerpt of non-news contents	6
Sponsor		Sponsoring message(s)	62
Newsbrief		Full news brief	18
Debate		Studio conversation	11
Story		Edited news story	103
Shot		News shot(s)	4
Clip		Musical clip, song...	9
Interlude		Interlude	4
Mix		Mix of several of the above	9
Error		Wrong detection	0

F. *Semi-supervised Labelling*

Navigating the database enables one to discover very different cases of repeats, but it becomes quickly cumbersome to keep track of all the findings. Therefore we added a possibility to label the discoveries: selecting a repeat, the user verifies the start and end times, and adds a label, specifying the type and a short free text description for the label. After confirmation, an extensive query through the database propagates the label to all the repeated segments, and a visual coloured cue that a label already exists, clickable for further inquiry, becomes visible on all affected timelines. Types and colour codes and numbers of collected labels are given in Table I. This effort only covers a small part of the database, but it is relatively easy to label a high number of recurrent events; these can be retrieved and further explored for later use.

VI. PERFORMANCES

A. *Fingerprint Robustness Measurements*

The robustness is measured on a 10-minute half-resolution (448x256) h264 8bits colour video+audio sequence. The sequence is distorted (Table II), and re-encoded in MPEG-1 at 200kbps. Distorted sequences are searched against the original sequence. The temporal recall measures both the missed detections and the underestimated duration; it is estimated dividing the summed detected durations by 600 seconds. Results do not change (same temporal recall, 100% precision) when adding 10-khours fingerprints to the *reference* base.

For compactness, are only listed in table II the distortion parameters that give a temporal recall of 98%, 80%, and 50%. Due to some long shots with very low activity in the test sequence, the most challenging test for the video fingerprint is splitting into random shots. The performances against more complex distortions are not measured, but we have found that the video fingerprint is generally robust to compositing, as in Fig. 4, provided that the area covered by the original picture area is active and large enough.

TABLE II. ROBUSTNESS CRITERIA TABLE

Distortion	**Distortion to obtain a temporal recall of:**		
	98%	***80%***	***50%***
Blur (box)	3x3	6x6	8x8
Uniform Noise	[-18,+18]	[-34,+34]	[-49,+49]
Gaussian Noise	31 dB PSNR	23 dB PSNR	20 dB PSNR
Zoom in & crop	110%	125%	146%
Zoom out	84%	50%	44%
H stretch & crop	123%	133%	146%
V stretch & crop	109%	127%	148%
Rotate & crop	53°, 137°	83°, 97°	never (min 53% at 90° & 270°)
Resolution (XxY)	308x176	228x130	< 84x48
Contrast -	83%	65%	38%
Contrast + & clip	133%	146%	172%
Other	Always better than 98%, on changes in hue, saturation, H or V flip, negate, monochrome, audio gain…		
Split into random shots (seconds)	24 sec.(audio) 60 sec. (video)	8 sec. (audio) 40 sec. (video)	6 sec. (audio) 16 sec. (video)
Low pass (audio)	1000Hz cut-off	725Hz cut-off	375Hz cut-off

In our experiments, the audio fingerprint usually gives results that have the same duration as the video fingerprint results, or up to 2% better. The main challenge for the lightweight audio fingerprints appears when two audio tracks are mixed (voice-over, added soundtrack): the temporal recall then quickly drops below 50%.

B. *Fingerprinting Speed Measurements*

All the fingerprints are computed on an 8-core Dell

PowerEdge R610 2x Xeon L5520, acquired in 2009. Computing the video or audio fingerprint of a half-resolution (488x256) compressed video file is on average 30 times faster than real-time, including de-compression, using only one core. At 80% load, the 8-core system currently downloads, computes audio + video fingerprints, and highly compressed copies of 24 1-hour files every day for 14 TV channels, i.e. 336 hours per day.

Fig. 4. An example of a correct detection, even though picture is zoomed out, re-framed, with a second video present.

C. Search Speed

For practical reasons, we run the searches through the *fingerprints bases* on a second, similar machine. Searching one 288h *daily fingerprints base* for repeats takes 7 minutes in video on 1 core. Searching the *twice trimmed daily fingerprints base* (~100 hours) against the *big base* (~180,000 hours) takes 2.2 hours on 3 cores. Assembling and indexing all the required fingerprints bases, searching, and voting, use on average 25% of the available power on the second machine. We also tested searching the *big base* for segments of video of various lengths. The results (all on 1 core, except for 100h) are summarised in Table III.

TABLE III. SEARCH SPEED (VIDEO)

Reference duration	**Candidate duration**						
	2 m	***10 m***	***1 h***	***24 h***	***40 h***	***100 h***	***288 h***
180,000 h	6 s	19 s	174 s	1h24	2h14	2h14 (3 cores)	--
288 h	0.05 s	0.06 s	0.14 s	2.57 s	--	--	0h07

VII. FINDINGS

Searching through this 1461-day database, we have discovered a number of interesting facts, and sometimes unexpected behaviours.

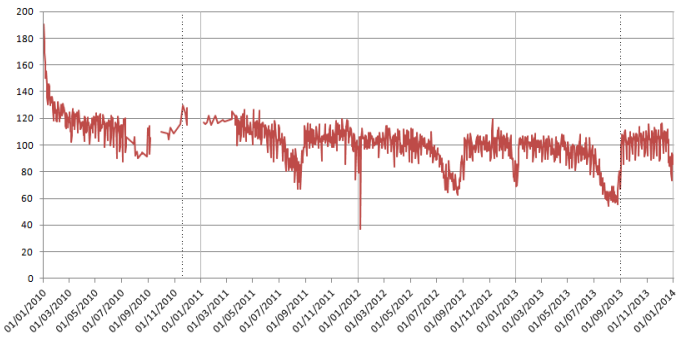

Fig. 5. Daily size increase, in hours per day, of the video *big base* over the period. Dashed lines mark the addition of new channels in Nov. 2010 and Sept. 2013. Low peak on 6/01/2012 is due to missing recorded files.

A. Fingerprints Database Size

The *big base* has reached 13 GB after four years (18 GB in audio), including only *fresh contents*' fingerprints. It would have exceeded 30 GB otherwise. Fig. 5 shows the increase, in hours per day, of the *big base* size. After a sharp initial fall, daily increase presents a slow downwards trend, despite the inclusion of two new channels: this is due to some long TV programmes (movies and documentaries) being re-broadcast after several years. During the winter and summer breaks, daily increase drops, due to a higher use of re-broadcasting.

B. Large Numbers

One 18-second title sequence for a daily game was found repeated on 1 channel on 1363 different days. According to the *electronic programme guide*, during the period 1366 such games were run, but we verified that the 3 missing ones had been cancelled due to political or sportive events.

Fig. 6. Snapshot of a title sequence found on 1363 different days

We found at least 8 sequences repeated across the whole set of channels: 6 were commercials, 2 were messages from the French TV regulating agency (CSA). The segment most repeated within one day (a commercial) was broadcast 125 times on 9 channels on 7/02/2012, 115 times on 8 channels on 10/10/2012, and less on 4 other days. A 29-second interlude was repeated (with slight changes) 181 times on 30/10/2011 on one channel during 1.5 hour.

TABLE IV. THE ANALYSED CHANNELS

Channel	***Contents***	***Average hours per day found***		
		R. same day	***Repeated***	***Fresh***
TF1	Major private channel	4.0 h	17.4 h	8.6 h
FR2	Major public channel	2.8 h	13.0 h	14.2 h
FR3	Major public channel	3.2 h	17.4 h	10.8 h
FR4	Public, for young people	2.7 h	22.8 h	2.3 h
FR5	Public cultural	3.4 h	20.5 h	7.0 h
M6	Private music and series	5.5 h	18.4 h	7.4 h
Arte	Cultural channel	2.0 h	20.6 h	7.3 h
Canal+	Major private paying	2.1 h	17.6 h	9.6 h
NRJ12	Private music and series	7.5 h	22.1 h	3.8 h
France 24	Public continuous news	18.1 h	20.6 h	7.4 h
I-Télé	Private continuous news	18.1 h	19.4 h	10.9 h
LCI	Private continuous news	14.8 h	15.6 h	9.4 h

C. Measurements and Trends

The collected TV channels are listed in Table IV. A number of trends can be observed, e.g. on the multiple timeline view in Fig. 2, and in Table IV. The major channels broadcast on average more than 8 hours of fresh contents per day, whereas lower-budget channels broadcast only 2 to 7 fresh hours. The continuous information channels re-use a considerable part of their contents: not only news stories, but also anchorperson

shots and debates; unlike the other channels, their re-use is concentrated on the same day, with very few repeats afterwards (mostly commercials); repeats are much shorter, except off-hours between 0:00 and 6:00, where 15-min recorded newscasts are played in loop, as visible on the 3 bottom timelines of Fig. 2.

Inter-programmes are often very visible on the timelines, as typical sequences: teasers/commercials/other/commercials/ /teasers, e.g. Fig. 3; the duration (3 to 15 minutes) can be measured on the timelines. Inter-programmes usually don't appear on the *electronic programme guides*; the programmes themselves start later than announced, and run for a shorter duration. Commercials are, by far, the most repeated segments, followed by jingles, credits and title sequences. Longer repeats are fewer, but account for the largest duration of the repeated contents on educational, music, and series channels.

VIII. Conclusion and Future Work

We hope to have shown that a global carefully designed, evenly distributed, lightweight fingerprint, applicable both to audio and to video, could be sufficient to detect, on a relatively affordable system, the repeated contents for more than 4 years and 10 TV channels. Some challenges remain: in video, duration is still under-estimated, contents with no activity are not detected; stretching or reversing the direction of time cannot be detected without changing the search and vote strategy. Being purely based on measured activity, such a fingerprint may appear to be fragile, but we have found in practice no evidence of false detection over our 380,000 hours experiment. The recall on shortest detections (<16s) would benefit from using both audio and video results to drive a second more sensitive *search* & *vote* pass. Searching through the mySQL results database was made possible through customised phpMyEdit search forms and timeline views. Some of the findings however required searching through the database directly using SQL queries.

We intend to keep this experimentation running, and to extend the number of channels. We have started to run the same experiment over a set of 25 regional TV channels with similar contents. Beyond numbers, we intend to improve the accessibility to the results. At the moment, the audio and video results are stored in two distinct databases, but little has been done to exploit the very high similarity between the two databases, or the differences that do exist.

Another track we wish to follow is to automate the generation of structured decomposition of the TV channels streams. Using the results of detections, and the data from the TV *electronic programme guides*, we would like to generate an accurate timing of the starting and ending of programmes, and of the - usually undocumented - inter-programmes. Authors such as Benezeth [15], Manson [16], Abduraman [17], Wu [18], Gauch [19], have undertaken such work with promising results, but we estimate that the reliability and scale of the obtained data should substantially help this work.

Acknowledgements

The authors wish to thank Rakia Jaziri, Elisabeth Chapalain, Marc Tarin, Paul Tomi, for their participation in the project, and Frédéric Dumas and Florent Lioret for their useful assistance and advices.

References

[1] A. Wang, "The Shazam music recognition service", Communications of the ACM - Music information retrieval, August 2006, Vol. 49 Issue 8, pp 44-48

[2] M. Yeh, K. Cheng "A compact, effective descriptor for video copy detection", Proceedings of the 17th ACM international conference on Multimedia 2009, pp 633-636

[3] B. Liu, Z. Li, Y. L. Yang, M. Wang, X. Tian. "Real-time video copy-location detection in large-scale repositories." Proc. ACM Multimedia 2011, vol. 18, no. 3, pp. 22–31

[4] K.M. Pua, J.M. Gauch, S.E. Gauch, J.Z. Miadowicz "Real time repeated video sequence identification" Computer Vision and Image Understanding 2004, vol. 93, no 3, pp 310-327.

[5] I. Laptev, T. Lindeberg. "Space-time interest Points". International Conference on Computer Vision 2003, pp 432–439,.

[6] A. Joly, O. Buisson, C. Frelicot. "Content-based copy detection using distortion-based probabilistic similarity search". IEEE Transactions on Multimedia, 2007, pp 293-306.

[7] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa. "Video copy detection: a comparative study". Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR2007, pp. 371-378.

[8] S. Jiang, L. Su, Q. Huang, P. Cui, Z. Wu, "A Rotation Invariant Descriptor for Robust Video Copy Detection", The Era of Interactive Media, 2013, pp 557-567

[9] L. Shang, L. Yang, F. Wang, KP. Chan, "Real-time large scale near-duplicate web video retrieval", Proceedings of the 2010 international conference on Multimedia, pp 531-540

[10] C. Herley, C. "ARGOS: automatically extracting repeating objects from multimedia streams", IEEE Transactions on Multimedia, 2006, vol 8, no1, pp115-129.

[11] M. Covell, S. Baluja, M. Fink, "Detecting Ads in Video Streams Using Acoustic and Visual Cues", Computer 2006, Vol. 39, no 12, pp 135-137

[12] S. Poullot, O. Buisson, M. Crucianu, "Scaling content-based video copy detection to very large databases", Multimedia Tools and Applications, 2010, Volume 47, Number 2, pp 279-306

[13] A. Joly, O. Buisson, "A posteriori multi-probe locality sensitive hashing", Proceedings of the 16th ACM international conference on Multimedia, 2008, pp 209-218

[14] J. Yuan, G. Gravier, S. Campion, X. Li, H. Jégou, "Efficient mining of repetitions in large-scale TV streams with product quantization hashing", ECCV Workshop on Web-scale Vision and Social Media, 2012.

[15] Y. Benezeth, S-A Berrani, "Unsupervised Credit Detection in TV Broadcast Streams", International Symposium on Multimedia (ISM), 2010 IEEE, pp 175-182

[16] G. Manson, S-A. Berrani, "Repetition density-based approach for TV program extraction." In Image Analysis for Multimedia Interactive Services, 2009. WIAMIS'09. 10th Workshop on, pp 181-184.

[17] A. Abduraman, S-A Berrani, B. Merialdo, "An unsupervised approach for recurrent tv program structuring", Proc. of the European Interactive TV Conference 2011, pp 123-126

[18] X. Wu, S. I. Satoh, "Temporal recurrence hashing algorithm for mining commercials from multimedia streams", Proc. of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2324-2327

[19] J.M. Gauch, A. Shivadas, A. "Finding and identifying unknown commercials using repeated video sequence detection", in Computer Vision and Image Understanding 2006, vol. 103 no 1, pp 80-88