Video Copy Detection: a Comparative Study

Julien Law-To, Li Chen, Alexis Joly, Ivan Laptev, Olivier Buisson, Valérie Gouet-Brunet, Nozha Boujemaa, Fred Stentiford

Video Copy Detection: a Comparative Study

Julien Law-To
Institut National de l'Audiovisuel
Bry Sur Marne, France
jlawto@ina.fr

Li Chen
UCL Adastral Park Campus
Martlesham Heath, Ipswich, UK
l.chen@adastral.ucl.ac.uk

Alexis Joly
INRIA, IMEDIA Team
Rocquencourt, France
alexis.joly@inria.fr

Ivan Laptev
INRIA, VISTA Team
Rennes, France
ivan.laptev@inria.fr

Olivier Buisson
Institut National de l'Audiovisuel
Bry Sur Marne, France
obuisson@ina.fr

Valerie Gouet-Brunet
INRIA, IMEDIA Team
Rocquencourt, France
valerie.gouet@inria.fr

Nozha Boujemaa
INRIA, IMEDIA Team
Rocquencourt, France
nozha.boujemaa@inria.fr

Fred Stentiford
UCL Adastral Park Campus
Martlesham Heath, Ipswich, UK
f.stentiford@adastral.ucl.ac.uk

ABSTRACT

This paper presents a comparative study of methods for video copy detection. Different state-of-the-art techniques, using various kinds of descriptors and voting functions, are described: global video descriptors, based on spatial and temporal features; local descriptors based on spatial, temporal as well as spatio-temporal information. Robust voting functions is adapted to these techniques to enhance their performance and to compare them. Then, a dedicated framework for evaluating these systems is proposed. All the techniques are tested and compared within the same framework, by evaluating their robustness under single and mixed image transformations, as well as for different lengths of video segments. We discuss the performance of each approach according to the transformations and the applications considered. Local methods demonstrate their superior performance over the global ones, when detecting video copies subjected to various transformations.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Algorithms

Keywords

Content-Based Video Copy Detection

1. INTRODUCTION

Growing broadcasting of digital video content on different media brings the search of copies in large video databases to a new critical issue. Digital videos can be found on TV Channels, Web-TV, Video Blogs and the public Video Web servers. The massive capacity of these sources makes the tracing of video content into a very hard problem for video professionals. At the same time, controlling the copyright of the huge number of videos uploaded everyday is a critical challenge for the owner of the popular video web servers. Content Based Copy Detection (CBCD) presents an alternative to the watermarking approach to identify video sequences and to solve this challenge.

The Robustness issue:
Two videos which are copies
Source video: *Système deux.* C. Fayard 1975 (c)INA

The Discriminability issue:
Two similar videos which are not copies (different ties)

Figure 1: Copy / similarity.

A crucial difficulty of CBCD concerns the fundamental difference between a copy and the notion of similarity encountered in Content-Based Video Retrieval (CBVR): a copy is not an identical or a near replicated video sequence but rather a transformed video sequence. Photometric or geometric transformations (gamma and contrast transformations, overlay, shift, etc) can greatly modify the signal, and therefore copies can in fact be visually less similar than other kinds of videos that might be considered similar. CBVR applications aim to find similar videos in the same visual category, like for example soccer games or episodes of soaps, but most of these detections would clearly represent false alarms in a CBCD application. Figure 1 illustrates the differences between CBCD and CBVR on two examples of a very similar non-copy video pair and a less similar pair of video copies. As already mentioned by X. Fang et. al.in

[6], another strong difference between CBCD and CBVR is the fact that, in a CBCD application, a query video can be infinite stream with no specific boundaries and only a small part of the query video can be a copy.

2. STATE OF THE ART

For the protection of copyright, watermarking and CBCD are two different approaches. Watermarking inserts non visible information into the media which can be used to establish ownership. However, very robust watermarking algorithms are not yet available. In a CBCD approach, the watermark is the media itself. Existing methods for CBCD usually extract a small number of pertinent features (called signatures or fingerprints) from images or a video stream and then match them with the database according to a dedicated voting function. Several kinds of techniques have been proposed in the literature: in [11] in order to find pirate videos on the Internet, Indyk et al. use temporal fingerprints based on the shot boundaries of a video sequence. This technique can be efficient for finding a full movie, but may not work well for short episodes with a few shot boundaries. Oostveen et al. in [20] present the concept of video fingerprinting or hash function as a tool for video identification. They have proposed a spatio-temporal fingerprint based on the differential of luminance of partitioned grids in spatial and temporal regions. B. Coskun et al. in [4] propose two robust hash algorithms for videos both based on the Discrete Cosine Transform (DCT) for identification of copies.

Hampapur and Bolle in [8] compare global descriptions of the video based on motion, color and spatio-temporal distribution of intensities. This ordinal measure was originally proposed by Bhat and Nayar [2] for computing image correspondences, and adapted by Mohan in [19] for video purposes. Different studies use this ordinal measure [10, 15] and it has been proved to be robust to different resolutions, illumination shifts and display formats. Other approaches focus on exact copy detection for monitoring commercials; an example being Y. Li et al. in [18] who use a compact binary signature involving color histograms. The drawback of the ordinal measure is its lack of robustness as regards logo insertion, shifting or cropping, which are very frequent transformations in TV post-production. In [12], the authors show that using local descriptors is better than ordinal measure for video identification when captions are inserted.

When considering post-production transformations of different kinds, signatures based on *points of interest* have demonstrated their effectiveness for retrieving video sequences in very large video databases, like in the approach proposed in [14] and described in section 3.2. Using such primitives is mainly motivated by the observation that they provide a compact representation of the image content while limiting the correlation and redundancy between the features.

3. TECHNIQUES COMPARED

This section presents the different techniques used for the comparative evaluation in this paper. They use global descriptors of video like the ones describe in section 3.1 with techniques based on the temporal activity, on the spatial distribution and on a spatio-temporal distribution. Section 3.2 presents techniques using local descriptions of the content. Section 3.3 defines the concept of a voting function and adapts it to different video descriptors.

3.1 Global descriptors

[Temporal]. This method[1] defines a global temporal activity $a(t)$ depending of the intensity I of each pixel (N is the number of pixels for each image).

$$a(t) = \sum_{i=1}^{N} K(i)(I(i,t) - I(i,t-1))^2$$

where $K(i)$ is a weight function to enhance the importance of the central pixels. A signature is computed around each maxima of the temporal activity $a(t)$. The spectral analysis by a classical FFT, lead to a 16-dimensional vector based on the phase of the activity.

[Ordinal Measurement]. The ordinal intensity signature consists in partitioning the image into N blocks; these blocks are sorted using their average gray level and the signature $S(t)$ uses the rank r_i of each block i.

$$S(t) = (r_1, r_2, ..., r_N)$$

The distance $D(t)$ is defined for computing the similarity of two videos (a reference R and a candidate C) at a time code t where T is the length of the considered segment.

$$D(t) = \frac{1}{T} \sum_{i=t-T/2}^{t+T/2} |R(i) - C(i)|$$

Hampapur and Bolle in [8] have described and tested this technique and they have shown that it has superior performances when compared to motion and color features. Li Chen has re-developped this technique for this study.

[Temporal Ordinal Measurement]. Instead of using the rank of the regions in the image, the method proposed by L. Chen and F. Stentiford in [3] use the rank of regions along the time. If each frame is divided in K blocks and if λ^k is the ordinal measure of the region k in a temporal window with the length M, the dissimilarity D between a query video V_q and a reference video V_r at the time code t is:

$$D(V_q, V_r^p) = \frac{1}{K} \sum_{k=1}^{K} d^p(\lambda_q^k, \lambda_r^k)$$

where:

$$d^p(\lambda_q^k, \lambda_r^k) = \frac{1}{C_M} \sum_{i=1}^{M} |\lambda_q^k(i) - \lambda_r^k(p+i-1)|$$

p is the temporal shift tested and C_M is a normalizing factor. The best temporal shift p is selected.

3.2 Local descriptors

[AJ]. This technique described in [13] by A. Joly st. al. is based on an improved version of the Harris interest point detector [9] and a differential description of the local region around each interest point. To increase the compression, the features are not extracted in every frame of the video but only in key-frames corresponding to extrema of the global intensity of motion [5]. The resulting local features are 20-dimensional vectors in $[0, 255]^{D=20}$ and the mean rate is about 17 local features per second of video (1000 hours of

[1]Thanks to G. Daigneault from INA who has developed this fingerprint.

video are represented by about 60 million feature vectors). Let $\vec{S}$ be one of the local features, defined as:

$$\vec{S} = \left(\frac{\vec{s_1}}{\|\vec{s_1}\|}, \frac{\vec{s_2}}{\|\vec{s_2}\|}, \frac{\vec{s_3}}{\|\vec{s_3}\|}, \frac{\vec{s_4}}{\|\vec{s_4}\|} \right)$$

where $\vec{s_i}$ correspond to 5-dimensional sub-vectors computed at four different spatio-temporal positions distributed around the interest point. Each $\vec{s_i}$ is the differential decomposition of the gray level 2D signal $\vec{I}(x, y)$ up to the second order:

$$\vec{s_i} = \left(\frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial y}, \frac{\partial^2 \vec{I}}{\partial x \partial y}, \frac{\partial^2 \vec{I}}{\partial x^2}, \frac{\partial^2 \vec{I}}{\partial y^2} \right)$$

[**ViCopT**]. **ViCopT** for video Copy Tracking is a system developed for finding copies. The system is fully described in [17]. Harris points of interest are extracted on every frame and a signal description similar to the one used in [AJ] is computed, leading to 20-dimensional signatures. The difference is that the differential decomposition of the gray level signal until order 2 is computed around 4 *spatial* positions around the interest point (in the same frame). These points of interest are associated from frame to frame to build trajectories with an algorithm similar to the KLT [22]. For each trajectory, the signal description finally kept is the average of each component of the local description. By using the properties of the built trajectories, a label of behavior can be assigned to the corresponding local description. For CBCD two particular labels have been selected:

- label *Background*: motionless and persistent points along frames
- label *Motion*: moving and persistent points

The final signature for each trajectory is composed of 20-dimensional vector, trajectory properties and a label of behavior.

[**Space Time Interest Points (STIP)**]. This technique was developed by I. Laptev and T. Lindeberg in order to detect spatio-temporal events (see [16]). The space time interest points correspond to points where the image values have significant local variation in both space and time. Previous applications of this detector concerned classification of human actions and detection of periodic motion. In this paper we are interested in the application of this detector to CBCD. For now, the detector has not been optimized for the task of copy detection and the presented results are preliminary. Space time interest points are described by the spatio temporal third order local jet leading to a 34-dimensional vector

$$j = (L_x, L_y, L_t, L_{xx}, ..., L_{ttt})$$

where L_{x^m, y^n, t^k} are spatio-temporal Gaussian derivatives normalized by the spatial detection scale σ and the temporal detection scale τ.

$$L_{x^m, y^n, t^k} = \sigma^{m+n} \tau^k (\partial_{x^m y^n t^k} g) * f$$

The L_2 distance is used as a metric for the local signatures.

3.3 Voting functions

A candidate video sequence is defined as a set of n_f successive frames described by features (global features or n_c local features). The features are searched in the database: a sequential search is performed for techniques with global features and an approximative search described in [13]. The similarity search technique provides for each of the n_c candidate local features a set of matches characterized by an identifier V_h defining the referenced video clip to which the feature belongs and their temporal position for the techniques described in section 3.1 and a spatio-temporal position for the techniques described in 3.2.

Once the similarity search has been done and once some potential candidates from the database have been selected, the partial results must be post-processed to compute a global similarity measure and to decide if the more similar documents are copies of the candidate document. Usually, this step is performed by a vote on the document identifier provided with each retrieved local feature [21]. In this study, we use a geometrically consistent matching algorithm which consists in keeping, for each retrieved document, only the matches which are geometrically-consistent with a global transform model. The vote is then applied by counting only the matches that respect the model (registration + vote strategy). The choice of the model characterizes the tolerated transformations. A generic model, if we consider resize, rotation and translation for the spatial transformations and also slow/fast motion from the temporal point of view, should be:

$$\begin{pmatrix} x' \\ y' \\ t'_c \end{pmatrix} = \begin{pmatrix} r\ cos\theta & -r\ sin\theta & 0 \\ r\ sin\theta & r\ cos\theta & 0 \\ 0 & 0 & a_t \end{pmatrix} \begin{pmatrix} x \\ y \\ t_c \end{pmatrix} + \begin{pmatrix} b_x \\ b_y \\ b_t \end{pmatrix} \quad (1)$$

where (x', y', t'_c) and (x, y, t_c) are the spatio-temporal coordinates of two matching points.

The transformation model parameters are estimated for each retrieved video clip V_h using the random sample consensus (RANSAC [7]) algorithm. Once the transformation model has been estimated, the final similarity measure $m(V_h)$ related to a retrieved video clip V_h consists in counting the number of matching points that respects the model according to a small temporal (for global features) and to a spatio-temporal precision for local features.

For **Temporal**, **Ordinal Measure** and **Temporal Ordinal Measure**, we have used a temporal registration without considereing slow/fast motion of the videos. For **AJ**, two different strategies taking into account the relative positions of the points are used in this study: a purely temporal method called **AJ_Temp** and a spatio temporal method called **AJ_Spatio_Temporal**. **STIP** has been associated to a temporal registration.

A special strategy is used to merge the local features of **ViCopT** because this system is asymmetric. The queries are points of interest while in the database, the features are trajectories. The registration is therefore finding the spatio-temporal offset that maximize the number of query points in the trajectories and is fully explained in [17]. It is close to the model used by **AJ_Spatio_Temporal**.

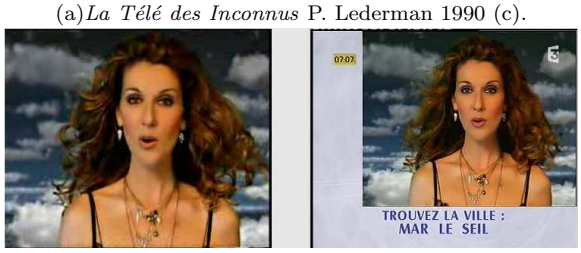
4. FRAMEWORK OF THE EVALUATION

Evaluating and comparing different systems of video copy detection is not obvious. This section presents the framework used for this study. A good CBCD system should have high precision (low false positive rate) and should detect all copies in a video stream possibly subjected to complex transformations.

4.1 Image transformations

CBCD systems should be able to cope with transformations of the post production process: insertion of logo, crop, shift. However, relevant transformations can also originate from involuntary events (re-encoding, cam-cording) which decrease the quality of the video and add blur and noise. Examples of transformations are shown in figure 2 with examples of detection found on the Internet by **ViCopT**. The example in Figure 2(a) shows a low quality re-encoded video effected by blur, noise and color changes. The example in Figure 2(b) presents a zoom with a shift and an insertion of a caption while Figure 2(c) demonstrates an example of vertical stretching. To test the robustness of the different CBCD methods, we simulate these types of transformations.

(a) *La Télé des Inconnus* P. Lederman 1990 (c).



(b) *Celine Dion Music Video* 2002 (c).

(c) *The Full Monty* 1997 (c) 20th Century Fox.

Figure 2: Copies found on the Web

4.2 Frameworks found in the literature

In the literature, evaluation usually consists in extracting a video segment from a video source. After different specific transformations, the video segment is used as a query. The goal is to find the best match and check if it is the right original video. For example, in [8], the authors use a 2 hrs 12 mins video reference (with a resolution 452 x 240) and video queries from 5.3 secs to 21.33 secs. Another evaluation of copy detection (see [18]) uses a database of 5000 videos with a constant length of 15 secs (total size is 21 hrs) and the queries are also videos with a constant length. These evaluation methods do not account for the fact that in a monitoring process, the queries are video streams without well-defined temporal boundaries. The fact that only a segment of the query video can be a segment of a video from the reference database has to be considered.

4.3 Proposed Benchmark

We created a set of video copies as follows: video segments were randomly extracted from a video database and transformed. These transformed segments were inserted in a video stream composed from videos not contained in the reference database. Figure 3 illustrate this benchmark. The query videos are analyzed using different methods under the goal of locating and identifying video segments corresponding to copies of the reference database. All the experiments

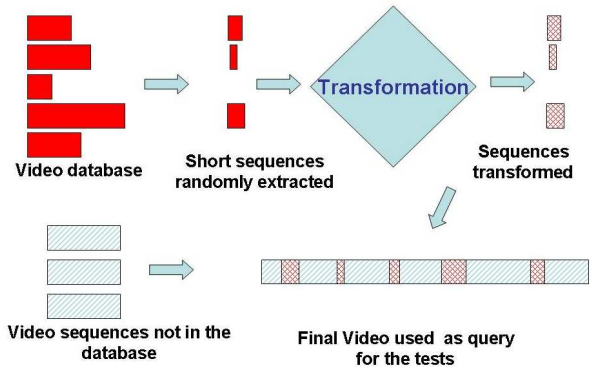


Figure 3: Video Benchmark Building.

were carried out on the BBC open news archives [1]. 79 videos (about 3.1 hours) cover different topics including conflicts and wars, disasters, personalities and leaders, politics, science and technology, and sports. The robustness to re-encoding is also tested, because to compute the query video, the video segment is re-encoded twice with different encoding systems.

4.4 Evaluation criteria

To evaluate our system, Precision Recall curves are computed using the following formulas:

$$Recall = \frac{N_{TruePositive}}{N_{AllTrue}}$$

$$Precision = \frac{N_{TruePositive}}{N_{AllPositive}}$$

Another criteria used is the average precision (AveP). The precision and recall are based on the whole list of detections returned by the system. Average precision emphasizes returning more relevant detection earlier. It is average of precisions computed after truncating the list after each of the relevant detections in turn:

$$AveP = \frac{1}{N_{max}} \sum_{r=1}^{N_{detected}} (P(r) \times \delta(r))$$

where r is the rank, N_{max} the number of relevant detections, $N_{detected}$ the number of detections, $\delta(r)$ a binary function on the relevance of a given rank and P(r) the precision at a given cut-off rank.

5. EXPERIMENTAL RESULTS

This section presents the results for the different CBCD methods on different test sets. Section 5.1 compares the different techniques with specific single transformations while section 5.2 uses a benchmark with mixed transformations.

5.1 Single transformation

For these experiments, 9 videos queries have been computed according to the framework defined in section 4.3: one with no transformation and 8 with a single transformation among the following:

- contrast increased by 25 %
- contrast decreased by 25 %
- crop 5% with black window
- gaussian radius-2 blur
- letter-box
- insertion of a small logo
- zoom 0.8 with black window
- zoom 1.2

Figure 4 illustrates these transformations. Each video query is 15 minutes long and contains 8 segments of simulated video copies (each segment 5 sec. long).

Figure 5 presents the PR curves for different techniques and for the different transformations. For all transformations, **Temporal Ordinal Measure** presents excellent results: all the segments have been found with no false alarm. The **Ordinal Measure** presents poor results for zooming, cropping and letter-box transformation.

Two reasons can explain a missing detection of a sequence. The first reason is that some descriptions are intrinsically not robust to some transformations. Technique based on Harris points of interest are not robust to a decrease of the contrast because the value of the corners can become too low. **Ordinal measurement** is by nature not robust to a crop which changes the average gray level of each block and thus spatial order sequence. Techniques which used a strict spatio temporal registration with no consideration of spatial deformation have poor results if we consider resize and letter-box transformations. The difference in these examples between **AJ_Temp** and **AJ_SpatioTemp** illustrates this decrease of quality because the spatial registration during the vote has eliminated some local descriptors well found during the search step.

The second reason of a missing detection is that the sequences themselves cannot be well described. If the query segments do not have much motion and activity, the temporal signature can not describe well this sequence. It is the same for **ViCopT**: there is a sequence with trajectories hard to compute and we can observe that the missing segment is always the same. For these techniques, the quality will also depend on the sequences and not only on the transformations.

5.2 Random transformations mixed

In real cases, there are combinations of transformations which can result on a large modification of the video. Figure 6 presents examples of real cases with large transformations found by **ViCopT** in a previous work. In the example (a), there is a large crop and captions inserted at the bottom but also on the persons. The example (b) shows a large insert of a logo and a zoom while the example (c) presents a change of background for creating a virtual duet for a TV show.

(a)Source video: *Alexandrie.* 1978 (c).

(b)Source video: *Samedi et Compagnie* 1970 (c)ORTF.

(c)Source video: *Gala du Midem.* G. Ulmer 1970 (c) INA

Figure 6: Complex transformation

To simulate this type of more complex transformation, we have created another video as defined in section 4.3. The length of each segment is a random value between 1s and 10s. For each segment, the transformations use different parameters and a random combination of transformations occurs. The shift and a change of the gamma had been added to the possible transformations but for this experiment, the zoom was not tested. There are 60 segments inserted in the video query which has a total length of 30 minutes. Examples of these simulated transformations are shown in figure 7. In the example (a), there is an horizontal and vertical shift, and insertion of a logo in the top-left corner and a change of the gamma.

Figure 8 presents the PR curves for the different techniques and table 1 gives the Average precision values. A system for controlling the copyright of a large number video files must have a great precision for avoiding too many false alarms that need to be checked by a person, so we compare the results at the precision 95%. At this precision, the best technique is **ViCopT** with a recall equal to 82%, then come **AJ_SpatioTemp** (recall 80%). The methods which use local features with only temporal registration has lower performances (55% for STIP and 51% for **AJ_Temp**). As we could expect for this test with large transformations and some short segments, the method with global features misses copies: **Temporal** has a 49% recall, **Temporal Ordinal Measure** has a 41% while the **Ordinal Measure** has a 10% recall.

The average precision values change a little bit this ranking and the **Temporal Ordinal Measure** and **AJ_Temp** perform better than STIP for this criteria.

Figure 4: Single transformations.

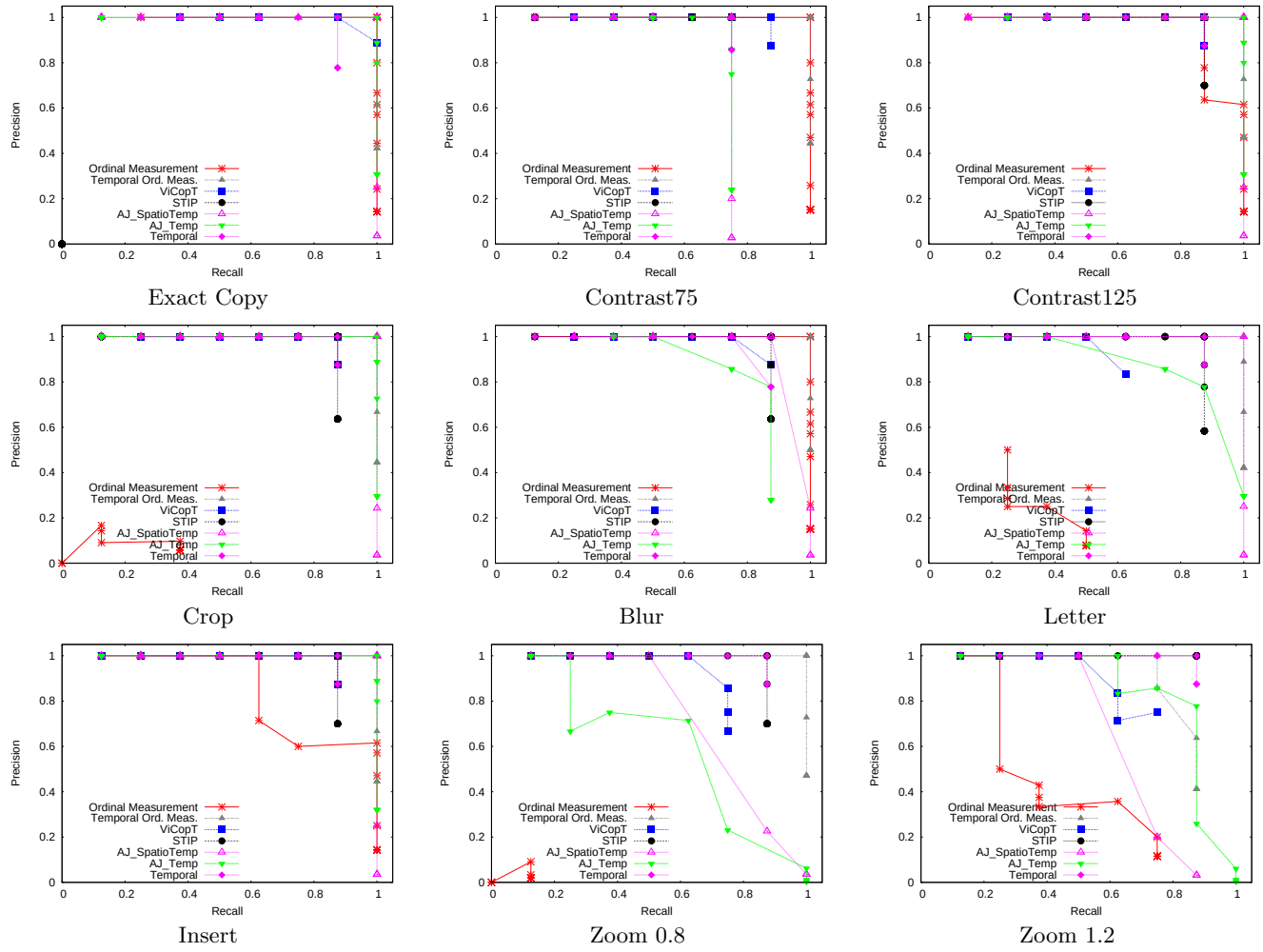
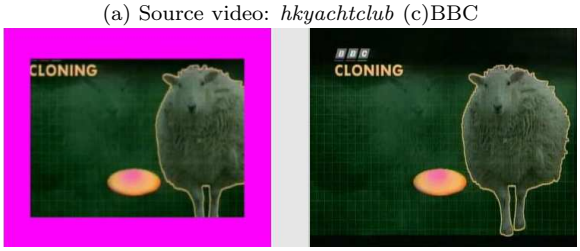


Figure 5: PR curves for single transformation.

(a) Source video: *hkyachtclub* (c)BBC



(b) Source video: *sheepclone* (c)BBC.

(c) Source video: *manhattanaftermath* (c)BBC.

Figure 7: Combined transformations.

Technique	AveP	Technique	AveP
ViCopT	0.86	STIP	0.55
AJ_SpatioTemp	0.79	Temporal	0.51
AJ_Temp	0.68	Ord. Meas.	0.36
Temp. Ord. Meas.	0.65		

Table 1: Average Precision For each technique.

6. DISCUSSIONS

This section presents some limits of this evaluation and give some reflexions about the advantages of the different techniques tested in this paper.

6.1 What technique should be used ?

How to choose a copy detection system will strongly depends on what we are searching for and where we are searching it. No single technique and no universal description seems to be optimal to all the applications that need video copy detection. We can identify different application cases for finding copies:

- Exact copies in a stream for statistics on commercials for example;
- Transformed full movie with no post production and possible decrease of quality (cam cording);
- Short segments on TV stream with possible large post production transformations;
- short videos on the Internet with various transformations (can be extract from TV stream);

For the first item, all the techniques should be efficient with a voting function that allows the detection to be precise

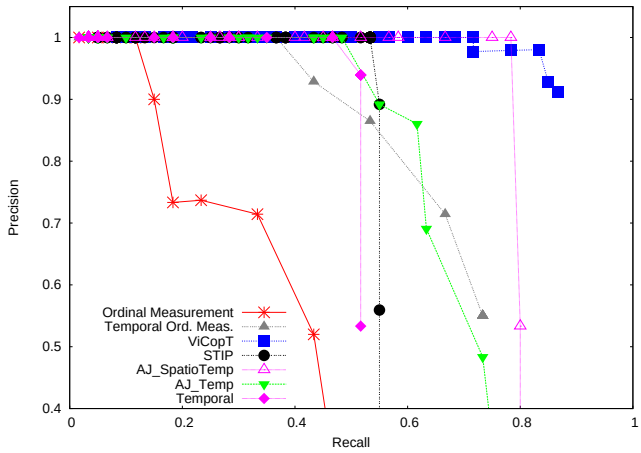


Figure 8: Video Benchmark Building.

for locating the boundaries. For finding full movies, as the length is important, methods with global features like *Temporal Ordinal Measurement* are probably faster with the same efficiency than methods with local features. The last two items are the most difficult cases because the segments can be short and strongly transformed. Finding short segments in a video stream is a critical issue for example, for the INA which provide video archives from a very large database (300 000 hours of digital videos) to TV channels. For this task **AJ_Temp** has proved a good efficiency and **ViCopT** presents a good improvement. For videos on the Internet, all the different difficulties are mixed and the solution depends on the quality required. The choice is still open but we think that local features seem more promising.

A limit of this evaluation is the size of the database (3 hours). The robustness issue has been evaluated on these experiments but the discriminability issue needs a larger database. Avoiding false positive detections is a very crucial point for an automatic or a semi automatic system. **AJ_Temp** has been tested on 30 000 hours in [13] with continuous TV stream as queries. It presents some false alarms when the similarity is too high: if the background is the same for a recurrent TV Show for example. **ViCopT** has been tested on 300 hours in [17] and seems to be more efficient with a better discriminability.

6.2 Some values

Less fundamental but practical reasons for using a particular method is its costs. This section gives some values related to these computational costs. The methods with global descriptions use a sequential search. It is fast but linearly dependent on the size of the database. The methods **ViCopT** and **AJ** use an index structure that allows the search to be very fast [14] and therefore this search is sub linear. Table 2 give some values measured during the tests. The number of features correspond to the features computed to describe the whole BBC archive database (3 hours of video). The search speed corresponds to the time needed to do the similarity search on a 15 minute query videos. As the tests were not done with the same computers and the same OS the results have to be analyzed with caution but the computers used were all classical PC. The

values for STIP are very high because no optimization and studies have been done for selecting the points and making the system efficient for an industrial use for now.

Technique	Nb of Features	Search Speed
ViCopT	218 030	27s
AJ	149 688	3s
Temp. Ord. Meas.	279 000	40min
STIP	2 264 994	3h30
Temporal	2700	34s
Ord. Meas.	279 000	37min

Table 2: Sizes of feature spaces and speed of the search.

7. CONCLUSIONS

In this paper, we have presented several solutions to the video copy detection issue, which is a very critical challenge for the protection of copyright. This work describes different techniques for characterizing the videos by their content. All the techniques have been tested and compared with a relevant dedicated evaluating framework, which is representative of real cases. We have first tested single transformations to measure the robustness of each technique. Then we have performed experiments on real cases where transformations are mixed and where copies can be very short segments. These experiments lead to the following conclusions: methods with local features have more computational costs but present interesting results in term of robustness. The **Ordinal Temporal measure** is very efficient for small transformations. However, additional tests need to be done. First, a bigger reference database should be used to test the discriminability of each technique. Secondly, a larger set of queries should be built for modeling other potential cases of use.

8. ACKNOWLEDGMENTS

This research has been conducted with the support of the European NoE MUSCLE http://www.muscle-noe.org/ .

9. REFERENCES

[1] http://www.bbc.co.uk/calc/news.

[2] D. Bhat and S. Nayar. Ordinal measures for image correspondence. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(4):415–423, 1998.

[3] L. Chen and F. W. M. Stentiford. Video sequence matching based on temporal ordinal measurement. Technical report no. 1, UCL Adastral, 2006.

[4] B. Coskun, B. Sankur, and N. Memon. Spatio-temporal transform-based video hashing. IEEE *Transactions on Multimedia*, 8(6):1190–1208, 2006.

[5] S. Eickeler and S. Müller. Content-based video indexing of tv broadcast news using hidden markov models. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 2997–3000, 1999.

[6] X. Fang, Q. Sun, and Q. Tian. Content-based video identification: a survey. In *Int. Conf. Information Technology: Research and Education*, 2003.

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[8] A. Hampapur and R. Bolle. Comparison of sequence matching techniques for video copy detection. In *Conference on Storage and Retrieval for Media Databases*, pages 194–201, 2002.

[9] C. Harris and M. Stevens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 153–158, 1988.

[10] X.-S. Hua, X. Chen, and H.-J. Zhang. Robust video signature based on ordinal measure. In *International Conference on Image Processing*, 2004.

[11] P. Indyk, G. Iyengar, and N. Shivakumar. Finding pirated video sequences on the internet. Technical report, Stanford University, 1999.

[12] K. Iwamoto, E. Kasutani, and A. Yamada. Image signature robust to caption superimposition for video sequence identification. In *International Conference on Image Processing*, 2006.

[13] A. Joly, O. Buisson, and C. Frelicot. Content-based copy detection using distortion-based probabilistic similarity search. IEEE *Transactions on Multimedia*, 2007.

[14] A. Joly, C. Frelicot, and O. Buisson. Feature statistical retrieval applied to content-based copy identification. In *International Conference on Image Processing*, 2004.

[15] C. Kim and B. Vasudev. Spatiotemporal sequence matching techniques for video copy detection. IEEE *Transactions on Circuits and Systems for Video Technology*, 1(15):127–132, Jan. 2005.

[16] I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, pages 432–439, 2003.

[17] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. In *ACM Multimedia, MM'06*, 2006.

[18] Y. Li, L. Jin, and X. Zhou. Video matching using binary signature. In *Int. Symposium on Intelligent Signal Processing and Communication Systems*, pages 317–320, 2005.

[19] R. Mohan. Video sequence matching. In *Int. Conference on Audio, Speech and Signal Processing*, 1998.

[20] J. Oostveen, T. Kalker, and J. Haitsma. Feature extraction and a database strategy for video fingerprinting. In *VISUAL '02: Proceedings of the 5th International Conference on Recent Advances in Visual Information Systems*, pages 117–128, London, UK, 2002. Springer-Verlag.

[21] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[22] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report CMU-CS-91-132, Apr. 1991.