VIDEO COPY DETECTION ON THE INTERNET:
THE CHALLENGES OF COPYRIGHT AND MULTIPLICITY

Julien Law-To ∗, Valerie Gouet-Brunet ○, Olivier Buisson ∗, Nozha Boujemaa ○

* Institut National de l’Audiovisuel (INA), Bry Sur Marne, France
○ Institut National de Recherche en Informatique et en Automatique (INRIA), Rocquencourt, France

ABSTRACT
This paper presents applications for dealing with videos on the web, using an efficient technique for video copy detection in large archives. Managing videos on the web is the source of two exciting challenges: the respect of the copyright and the linkage of multiple videos. We present a technique called ViCopT for Video Copy Tracking which is based on labels of behavior of local descriptors computed along video. The results obtained on large amount of data (270 hours of videos from the Internet) are very promising, even with a large video database (700 hours): ViCopT displays excellent robustness to various severe signal transformations, making it able to identify copies accurately from highly similar videos, as well as to link similar videos, in order to reduce redundancy or to gather the metadata associated. Finally, we also show that ViCopT goes further by detecting segments having the same background, with the aim of linking videos of the same category, like forecast weather programs or particular TV shows.

1. INTRODUCTION
Due to the increasing use of the World Wide Web and the wide-spread availability of ADSL, many web sites now propose space to store and broadcast personal video content on the web. As video archive professionals need to trace the uses of their videos and as video web servers need to automatically control the copyrights of the videos uploaded by the users, finding copies in a large video database has become a critical new issue. To identify video sequences, Content Based Copy Detection (CBCD) presents an alternative to the watermarking approach.

A crucial difficulty is the fundamental difference between a copy and the notion of similar image 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. These photometric or geometric transformations (gamma and contrast transformations, overlay, shift, etc) can greatly modify the signal. The system proposed here has two objectives: to identify the copies for the copyright issue and to link similar videos with the aim of reducing redundancy and of gathering the annotations associated. Figure 1 illustrates the two challenges with very similar videos but not copies and copies which are less similar.

Two similar videos which are not copies (different ties)

Two videos which are copies (one is used to make the other)

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

Fig. 1. Similarity / Copy.

2. RELATED WORK ON COPY DETECTION
CBCD generally consists in extracting a small number of pertinent features (called signatures or fingerprints) from the images or the video stream and matching them with the database according to a dedicated voting function. Several kinds of techniques have been proposed in the literature: in [1] in order to find pirated video on the Internet, Indyk et al. use temporal fingerprints based on the shot boundaries of a video sequence. Hampapur and Bolle in [2] compare global descriptions of the video based on motion, color and spatio-temporal distribution of intensities. They have shown that their signature, based on the ordinal signature, performs better than those based on motion and color. This ordinal measure was originally proposed by Bhat and Nayar [3] for computing image correspondences, and adapted by Mohan in [4] for video purposes. It consists in dividing the image into blocks; these blocks are sorted using their average gray level and the signature is the rank of each block. Different works use this ordi-
nal measure [5, 6] and it has proved to be robust to changes of the frame rate and resolution. 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. For example in [7], the authors show that using local descriptors is better than ordinal measure for video identification when captions are inserted. When considering these kinds of image transformations, signatures based on points of interest have demonstrated their effectiveness for retrieving video sequences in very large video databases, like the approach proposed in [8]. This method uses the temporal activity of the video in order to select some key frames and then extract Harris points of interest from these key frames. 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 detected features.

3. PROPOSED CONCEPT

We propose a solution that involves estimating and characterizing trajectories of points of interest throughout the video sequence. We plan on taking advantage of such trajectories to characterize the spatio-temporal content of videos: it allows the local description to be enriched by adding a spatial, dynamic and temporal behavior of this point. The aim is to provide a rich, compact and generic video content description where the behavior of a point along a trajectory can be seen as a temporal context of this point.

3.1. Video description

We present here a brief description of our indexing algorithm. Harris points of interest are extracted on every frame and a signal description is computed leading to the following 20-dimensional signatures: $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4)$ where the $\mathbf{s}_i$ are 5-dimensional sub-signatures computed at 4 spatial positions around the interest point. Each $\mathbf{s}_i$ is a differential decomposition of the gray level signal until order 2. Such a description is invariant to image translation and to affine illumination changes. These points of interest are associated from frame to frame to build trajectories with an algorithm similar to the KLT [9]. For each trajectory, the signal description finally kept is the average of each component of the local descriptors. The redundancy of the local description along the trajectory is efficiently summarized (the number of features is reduced by 50 in average in our experiments), with a reduced loss of information. By using the properties of the built trajectories, a label of behavior can be assigned to the corresponding local description: the categories of behaviors are simply obtained with heuristics and thresholds. This leads to different levels of description:

- Low-Level: spatio-temporal description of the signal.
- Mid-Level: trajectory parameters.
- High-Level: labels based on the behavior of points (temporal context).

The low and mid-level descriptors are obtained at the end of a purely bottom-up process, independent of the application: it is a generic description of the video, and is computed only once. The final high-level description follows from a top-down process, that is specific to the application.

3.2. On-line Retrieval for CBCD

As the off-line indexing part needs long time computational (1000 hours for dealing with 700 hours of video) and as the system of retrieval needs to be in real-time, the whole indexing process described in sections 3.1 can not be done for the candidate video sequences. A more fundamental reason is that the system has to be robust to small video insertion, or to re-authored video. The retrieval approach is therefore asymmetric and queries are local descriptors selected as following: every $p$ frames, $n$ points of interest are extracted. The advantage of the asymmetric technique is an on-line choice of the number of queries and the temporal precision, which gives flexibility to the system.

Fig. 2. Illustration of labels of behavior dedicated to Copy Detection. The boxes represent the amplitude of moving points along their trajectory (motionless points do not have a box). The “+” are the mean position of such points.

For CBCD, we select two particular labels: the motionless and persistent points along frames are used to define the label Background\(^1\) while the moving and persistent points define the label Motion\(^2\). These two labels, illustrated on figure 2, were chosen because it seems natural that they are useful for copy detection. The background is very robust and typical of several TV contents (talk shows for example) while motion brings distinctiveness to the video sequence. Moreover, selecting points according to these labels makes it possible to efficiently reduce the number of features (they are divided by 200 in our experiments).

Then, a voting function is necessary: it consists in exploiting the results of a statistical search which select some

\(^1\)amplitude of motion < 3 pixels and persistence > 15 frames
\(^2\)amplitude of motion > 10 pixels and persistence > 10 frames
candidates. In our approach, a spatio-temporal registration is done using the trajectory parameters of these candidates. This robust voting function is detailed and evaluated in [10] on synthetic and real cases.

It has been shown in [11] that this system is very efficient comparing to other approaches which use global features for describing the videos. It is fast: for treating 1 hour of video queries when considering a 700-hour video database as a reference, the system needs only 10 minutes to identify and to give the temporal boundaries of the copied segments.

4. EXPERIMENTS

We have already demonstrated the relevance and the efficiency of ViCopT on TV cases in [10] but on the Internet, the transformations are more various. As a lot of videos are taken from TV, modified and uploaded by users, post-production transformations occur but there is also a decrease in quality. For example, figure 3 presents detections obtained with ViCopT: in (a), the candidate video is blurred and presents a crop and a caption inserted while in (b) there is a large insert and a change of color; (c) presents a strong change in illumination which is non homogeneous due to a camcording system.

As a reference database, we have used 700 hours of videos chosen from the video archive database stored at INA (the French Institut National de l’Audiovisuel). These videos have a very diverse content from TV and were stored in MPEG-1 (25 fps). As video queries, we have downloaded for experimental purposes 21 videos that was related to INA database.

In order to compare the efficiency of ViCopT for videos on the Internet, the technique described in [8] was used as a reference. A strong difference is that we have enriched the local description with a temporal behavior of interest points. Moreover ViCopT describes the whole video sequence during the off-line indexing and not only the key images. This reference is used rather than [2] for example because the global description is not enough robust for large transformations such as those of figure 3.

Figure 4 presents the Precision/Recall curves obtained for the two techniques. The first curves consider the segments detected: they show a high precision for the two techniques but a recall better for ViCopT (+30%). The second ones are based on the frames detected, which correspond to the precision of the detected segments. This aspect is important for separating copies from videos having segments in common. The recall is also better (+45%). These experiments show the great performances of ViCopT for copy detection on the Internet.

5. APPLICATIONS

The applications presented here show the relevance of our technique facing two fundamental problems when considering videos on the Internet: the respect of the copyright and the management of multiplicity of videos.

5.1. Traceability and protection of digital content

As evaluated in section 4, the system developed can be an efficient tool for filtering the copyright content on the web. For experimental purposes, we have downloaded 5600 videos, for a total of 270 hours, on different web sites by using some relevant keywords related to the INA archives and those videos were used as queries. ViCopT found 196 videos with copyrighted contents from the INA archives (3.5% of the videos).

5.2. Web multiplicity and video linkage

As mentioned by S-C. Cheung and A. Zakhor in [12], many popular video content are being mirrored and republished re-
resulting in excessive content duplication. On the popular web sites that store videos, the number of videos is estimated to be equal to 50 millions of videos and this number is increased by 20% every month. To measure the multiplicity among those videos, we performed the following experiment: all the videos have been searched on themselves (270 hours of query facing the same 270-hour video database). Every video should be linked at least to itself and it is the case for 97% of the videos. This shows that our asymmetric system is efficient for finding copies with very different quality and sources. ViCopT can be used for eliminating the redundancy.

More than eliminating redundancy, linking videos in a large database is a very motivating challenge: the description can be enriched by the different descriptions added by users (tags, keywords) in a Web 2.0 philosophy\(^3\). Different types of links can be defined: two videos can be copies or can just contain common short segments (less than 10 seconds of common frames). Moreover, an advantage of ViCopT is that we have defined labels of behavior for each local descriptor and we can use specific labels to link videos. To illustrate this idea, we have used here the label Background. For evaluating the relevance of our approach, 30 videos of forecast weather programs with different speakers were added to the 5600 queries of section 5.1. The 30 videos were successfully linked between themselves among the 5630 videos with no false alarm. Figure 5 presents another example of similarities found by exploiting this label and table 1 sums up the results for the different types of linkage considered.

![Fig. 5. Similar videos found according to label Background. The videos retrieved belong to the same TV show.](image)

<table>
<thead>
<tr>
<th>Type of link</th>
<th>nb of linked videos</th>
<th>maximum nb of links</th>
<th>average nb of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copies</td>
<td>304</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Common parts</td>
<td>224</td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>Background</td>
<td>527</td>
<td>28</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1. Linking videos according to different criteria.

\(^3\)which refers to a supposed second generation of Internet-based services that emphasize online collaboration and sharing among users.

6. CONCLUSION

This paper presents different contributions dedicated to the new challenges of videos on the Internet. We have shown the relevance of our approach for finding copies in a large video database and therefore for filtering the copyright web contents. By searching the videos on themselves, we can create links for duplicate and copies and therefore avoid duplicates and merge the metadata of these videos. Last contribution is a different use of the labels defined during the indexing step for finding similarities due to the background and therefore we can link the videos purely by the content.

7. REFERENCES


