CONTENT-BASED VIDEO COPY DETECTION IN LARGE DATABASES: A LOCAL FINGERPRINTS STATISTICAL SIMILARITY SEARCH APPROACH

Alexis Joly[(1,2)], *Carl Frélicot*[(2)] *and Olivier Buisson*[(1)]

[(1)] Département Recherche et Études, Institut National de l'Audiovisuel, 94366 Bry/Marne, France
[(2)] Labo. d'Informatique-Image-Interaction, Université de La Rochelle, 17042 La Rochelle, France

ABSTRACT

Recent methods based on interest points and local fingerprints have been proposed to perform robust CBCD (content-based copy detection) of images and video. They include two steps: the search for similar local fingerprints in the database (DB) and a voting strategy that merges all the local results in order to perform a global decision. In most image or video retrieval systems, the search for similar features in the DB is performed by a geometrical query in a multidimensional index structure. Recently, the paradigm of approximate k-nearest neighbors query has shown that trading quality for time can be widely profitable in that context. In this paper, we evaluate a new approximate search paradigm, called *Statistical Similarity Search* (S^3) in a complete CBCD scheme based on video local fingerprints. Experimental results show that these statistical queries allow high performance gains compared to classical ϵ-range queries and that trading quality for time during the search does not degrade seriously the global robustness of the system, even with very large DBs including more than 20,000 hours of video.

1. INTRODUCTION

Content-Based Copy Detection (CBCD) schemes are an alternative to the watermarking approach for persistent identification of images and video clips [1, 2, 3, 4]. It generally consists in extracting as few features as possible from the candidate objects and matching them with a DB. Since the features must be discriminant enough to identify an image or a video, the features are often called *fingerprints* or *signatures* [4]. Recently, robust CBCD schemes based on local fingerprints have been proposed to deal with geometrical transformations, such as resizing, shifting or inserting [1, 3]. In these techniques, the detection includes two steps: the search for similar local fingerprints in the DB and a voting strategy that merges all the local results in order to decide which of these results are some copies of the candidate object. The method proposed in [1] is dedicated to static images and the voting strategy is only based on the image identifiers of the local fingerprints returned by the search. In [3], we proposed a method dedicated to video. The experimental results we present in this paper were obtained in that context.

As for many content based retrieval systems, one of the difficult task of a CBCD scheme is the time cost of the *similarity search* in the fingerprints reference DB, which can be very large. To solve this problem, most multimedia retrieval systems use multidimensional index structures [5, 6] or improved sequential techniques, e.g. the VA-file [7]. For the last few years, researchers are interested in trading quality for time [8, 9, 10, 11] and the paradigm of *approximate similarity search* has emerged [12]. Some of the proposed solutions are simply *early stopping approaches* [9] and other techniques, e.g. [13], are based on geometrical approximations during the filtering rules of the search algorithm. More recent techniques are based on a probabilistic selection of the bounding regions used in the indexing structure [10, 11]. They allow to control directly the expected percentage of the real k-nearest neighbors of the query. Range queries are rarely used because the results of the similarity search are almost always directly linked to the results that are provided to the user, for whom it is useful to get always the same number of results. We think however that a k-nearest neighbor search is not appropriate to copy detection and especially for techniques that include a voting strategy after the search. The main reason is that the number of relevant fingerprints for a given query is highly variable. In a large TV archives DB, several video clips can be duplicated 600 times, whereas other video clips are unique. Furthermore, inside a single video sequence, points of the background are detected many times whereas others corresponding to moving objects are unique.

The basic idea of the *Statistical Similarity Search* (S^3) technique we proposed in [14] was to extend the approximate search paradigm to ϵ-range queries, leading to the *statistical query* paradigm (section 2). In this paper, we evaluate the S^3 technique in a complete video CBCD scheme, described in section 3. The evaluation description and the results are provided in section 4.

2. STATISTICAL QUERY

By excluding several regions of an hyperspherical query, having a too small intersection with the bounding regions of the index structure, it is possible to obtain high speed-up with very small losses in the results. However, it is not possible to take the volume percentage as an error measure because it would be equivalent to consider that the relevant similar fingerprints are uniformly distributed inside an hypersphere. When the dimension increases, the fingerprints following such a distribution become closer and closer to the surface of the hypersphere but this is not true in reality, as illustrated on Fig. 1. The solid curve (left) is the real distribution of the distance between referenced and distorted fingerprints issued from a transformed version of the referenced video sequences for the same interest points. In this example, the transformation was a resize of factor $w_{sc} = 0.8$, but we observed the same kind of distributions for all studied transformations, including colorimetric distorsions and noise addition. The two other dotted curves represent the estimated probability density function for two probabilistic models: an uniform spherical distribution (right) that would be obtained if we took the volume percentage as an error measure and

a zero mean normal distribution (center) under components independence assumption. A simple independent normal distribution is much closer to the real distribution than the uniform one.
The proposed *statistical query* paradigm relies on the distribution of the relevant similar fingerprints. Let the *distorsion* vector ΔS be defined by:

$$\Delta S = S(m) - S(t(m)), \quad t \in T$$

where $S(m)$ is the fingerprint of a referenced pattern m, $S(t(m))$ is the distorted fingerprint, i.e the fingerprint of the transformed pattern $t(m)$ and T is the set of transformations that can be applied between a referenced sequence and a copy. The *statistical query of expectation* α is defined as the search of all the fingerprints contained in a region V_α of the feature space satisfying:

$$\int_{V_\alpha} p_{\Delta S}(X - Q)\, dX \geq \alpha \tag{1}$$

where Q is a candidate fingerprint and $p_{\Delta S}(.)$ is the probability density function of the distorsion.

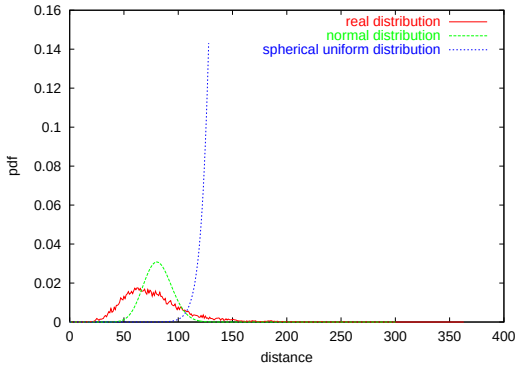


Fig. 1. Distribution of the distance between a fingerprint and its distorted version after resizing of a video sequence ($w_{sc} = 0.8$)

In practice, the first step of a search in a multidimensional indexing structure is a set of geometric filtering rules that quickly exclude most of the bounding regions. To process the statistical queries in an indexing structure, we propose to replace the geometric rules by probabilistic rules, according the distortion model. The main advantage is that a statistical query has no intrinsic shape constraint. Thus, the region V_α that makes equal the probability to find a relevant fingerprint to α is naturally adapted to the shape of the bounding regions used by an indexing structure. Details on the indexing structure and the search algorithms of the S^3 technique are provided in [14]. The distorsion model used in the rest of this paper is a zero mean normal distribution independent for each component:

$$p_{\Delta S j}(x_j) = f_{\mathcal{N}(0,\sigma)}(x_j)$$

where the unique parameter σ is estimated on a set of representative transformations.

3. VIDEO LOCAL FINGERPRINTS AND VOTING STRATEGY

The local fingerprints extraction [3] includes three steps:

- a key-frame detection, based on the mean of the frames difference also called *intensity of motion*.
- an interest point detection in each key-frame, processed by an improved version of the Harris detector [15].
- a local characterisation computed around each interest point, leading to a $D = 20$ -dimensional fingerprint S:

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

where the s_i are 5-dimensional sub-fingerprints computed at four different spatio-temporal positions distributed around the interest point. Each s_i is a differential decomposition of the graylevel 2D signal $I(x, y)$ until the second order:

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

In the indexing case, the fingerprints are simply inserted in the indexing structure with a video sequence identifier Id and a time-code tc. In the detection case, the statistical query returns for each candidate fingerprint S_j a set of K_j referenced fingerprints $\{S_{jk}\}_{k \in K_j}$ with their identifiers $\{Id_{jk}\}_{k \in K_j}$ and their time-codes $\{tc_{jk}\}_{k \in K_j}$. These results are stored in a buffer for a fixed number of key-frames in order to estimate the best sequences. We note N_{cand} the number of candidate fingerprints contained in the corresponding time interval ($j \in [1, N_{cand}]$).
The estimation is only based on the identifiers and the time-codes and not on the fingerprint itself. For each identifier id, the corresponding time-codes are used to estimate the unique parameter b of the following temporal model:

$$tc' = tc + b$$

where tc' represents a time-code of the candidate sequence and tc a time-code of a referenced sequence. This simple estimation problem is solved by the following minimization equation:

$$\widehat{b}(id) = \arg\min_b \left(\sum_{j=1}^{N_{cand}} \min_{\substack{k \in K_j \\ Id_{jk} = id}} \rho\left(\left| tc'_j - (tc_{jk} + b) \right| \right) \right) \tag{2}$$

where $\rho(u)$ is a non-decreasing cost function allowing to decrease the contribution of outliers. The Tukey's biweight M-estimator was chosen for $\rho(u)$ (see [16] for details). Once $\widehat{b}(id)$ has been estimated, a similarity measure n_{sim} is computed for each identifier id represented in the results by a voting strategy. It simply consists in counting the number of candidate fingerprints (i.e the number of interest points) that contribute to the solution $\widehat{b}(id)$ according to a small tolerance interval. By thresholding the value of n_{sim}, we finally decide which of the identifiers represented in the results correspond effectively to a copy of the candidate sequence.

4. EXPERIMENTS AND RESULTS

The DBs used in these experiments contain real video local fingerprints extracted by the method described in section 3. The referenced video sequences come from a french TV archives DB that contains all kinds of TV programs from the Fourties at our days: news, sport, show, variety, films, reports, black&white archives,

advertissements, etc. The average number of local fingerprints per hour of video is about $50,000$. Thus, a DB representing $10,000$ hours of video contains about $500,000,000$ fingerprints and the size of the corresponding DB file is about 13 Gb ($D = 20$ dimensional fingerprints + identifiers + time codes). Experiments were computed on a Pentium IV (CPU 2.5 GHz, cache size 512Kb, RAM 1.5 Gb).

4.1. Statistical query compared to exact range query

In this first experiment, we compare the search time of a statistical query to those of a classical spherical range query of radius ϵ (ϵ-range query). We do not aim at testing the relevance of the distorsion model, but at showing the advantage of a statistical query compared to an exact range query when the distribution is perfectly known.
We randomly select 1000 real fingerprints S in the DB and construct 1000 queries $Q = S + \Delta S$, where the components of the distorsion vector ΔS are independently generated according a zero mean normal distribution $p_{\Delta Sj}(x_j) = f_{\mathcal{N}(0,\sigma)}(x_j)$ with $\sigma = 18.0$. These queries are then searched in the DB using both a statistical query and an ϵ-range query. For different values of the query expectation α, we measure, for both query types, the average time of a single search (Fig. 2) and the retrieval rate (Fig. 3), i.e. the percentage of queries for which the original fingerprint S belongs to the results returned by the search. The radius ϵ of the range query was set in order to have the same expectation α than the statistical query. It is indeed easy to show that for the given distorsion model, the L2 norm of the distorsion has the following probability density function:

$$p_{\|\Delta S\|}(r) = \frac{f_{\mathcal{N}(0,\sigma)}(r)}{(2\pi\sigma)^{\frac{D-1}{2}}} \frac{\pi^{\frac{D}{2}} D}{\Gamma\left(\frac{D}{2}+1\right)} r^{D-1}$$

where Γ is the gamma function and D is the dimension of the feature space. By tabulating the values of the corresponding cumulated density function, it is easy to choose the value of the radius ϵ such as

$$\int_0^{\epsilon} p_{\|\Delta S\|}(r)\, dr = \alpha$$

The average search time curves of Fig. 2 (displayed in logarithmic coordinates) show that the statistical query approach outperforms the classical exact range query. Depending on α, it is from 17 to 132 times faster. The geometrical constraint of an exact ϵ-range query degrades seriously the search time without improving the retrieval rate, as shown on Fig. 3. This result does not depend on the index structure. Whatever the shapes of the bounding regions are, it is indeed well known that the number of intersections with an hypersphere becomes very high when the dimension increases. The main asset of the statistical query is that it does not impose any particular shape. It only uses the probability to find a relevant fingerprint inside the bounding regions.

4.2. Robustness of the video CBCD system

The purpose of these experiments is to study the interaction between the S^3 technique and the global robustness of the video CBCD system. We extract randomly 100 video sequences of 10 seconds each from the reference DBs and apply to them five kinds of transformations: vertical shift of w_{sh} $\%$ of the image, resize of factor w_{sc} and gamma modification ($I'(x,y) = I(x,y)^{w_\gamma}$).

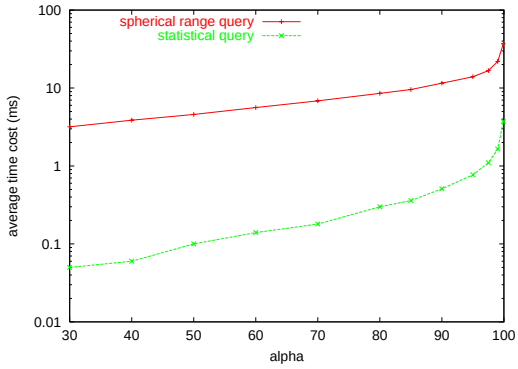


Fig. 2. Average search time (ms) $vs.$ query expectation α

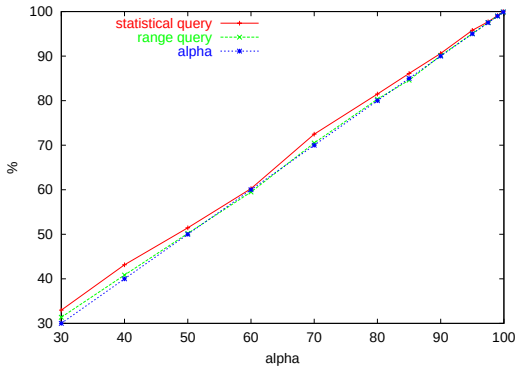


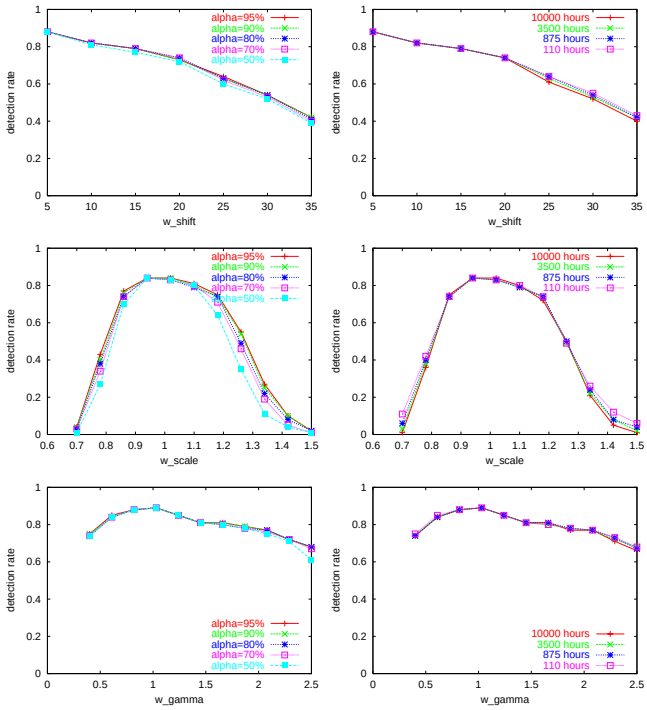
Fig. 3. Retrieval rate (%) $vs.$ query expectation α

The 100 transformed sequences are submitted as candidates to the video CBCD system and we measure the good detection rate with a tolerance of 2 frames. The threshold value of n_{sim} was set so that in average less than 1 false alarm occurs per hour when the system is continuously monitoring a TV channel. The parameter σ of the distorsion model was set to 20.0. Different DB sizes, different values of the expectation α and different values of tranformations parameters are used in these experiments. Results are presented on Fig. 4 in the form of abacuses of

- α (left-hand side) using a DB containing about 3500 hours of video
- the DB size (right-hand side) with α set to 80%

Two tables presenting the average search time of one single fingerprint for the different values of α and DB size are given at the bottom of the figure. Right subplots show that the DB size does not affect so much the detection rate whatever the transformation is. The main reason is that the statistical query guarantees the same expectation for the similarity search whatever the DB size is. The increased number of false retrieved fingerprints does not degrade the final quality thanks to the voting strategy which is highly discriminant (the temporal coherence of many fingerprints is very rare). It is important to note that it would not have not been the same if we had used a k-nearest neighbor search. When the DB size is multiplied by 100, the higher the density of fingerprints, the higher the chance to exclude relevant fingerprints from the results. Left subplots show that the detection rate remains almost invariant for all transformations as the expectation α decreases 95% downto 70% whereas the search is 4 times faster. For the most severe trans-

formations, it begins to fall down when α equals 50%. The most important result of this experiment is that an approximate search is particularly profitable when a voting strategy is employed after the search. It is indeed useless to retrieve the less distortion-invariant fingerprints since they seriously degrade the search time without really improving the robustness of the video CBCD system.



α (%)	time cost (ms)
50	1.05
70	2.15
80	3.45
90	5.76
95	8.64

video hours	time cost (ms)
110	0.69
875	1.81
3,500	3.45
10,000	9.11

Fig. 4. Detection rates and search times for several values of α (left) and several DB sizes(right)

5. CONCLUSION AND PERSPECTIVES

In this paper, we evaluated a new approximate search paradigm based on statistical queries. We showed that the geometrical constraint of a classical exact ϵ-range query can degrade seriously the time cost of the search without systematically improve the results. We showed that the use of a statistical filtering instead of an exact geometrical filtering can lead to a search about 100 times faster. We then studied the influence of our statistical similarity search strategy on the global robustness of a video CBCD system which is based on local fingerprints retrieval and a voting strategy. The experiments showed that in such a scheme, trading quality for time during the search is highly profitable, even when the size of the DB becomes very large.

The voting strategy will be our main future research topic for two reasons. Firstly, when the DBs becomes very large, the number of retrieved fingerprints during the search increases seriously and this step will probably become a new bottleneck of the total time cost of the CBCD system. Secondly, we would like to extend the estimation step on the time codes to the spatial positions of the interest points to improve the discriminance. This will however complicate the simple parameter estimation problem of the S^3 method and will require much more efficient algorithms.

6. REFERENCES

[1] S-A. Berrani, L. Amsaleg, and P. Gros, "Robust content-based image searches for copyright protection," in *Proc. of ACM Int. Workshop on Multimedia Databases*, 2003, pp. 70–77.

[2] A. Hampapur and R. Bolle, "Comparison of sequence matching techniques for video copy detection," in *Proc. of Conf. on Storage and Retrieval for Media Databases*, 2002, pp. 194–201.

[3] A. Joly, O. Buisson, and C. Frélicot, "Robust content-based video copy identification in a large reference database," in *Int. Conf. on Image and Video Retrieval*, 2003, pp. 414–424.

[4] J. Oostveen, T. Kalker, and J. Haitsma, "Feature extraction and a database strategy for video fingerprinting," in *Proc. of Int. Conf. on Visual Information and Information Systems*, 2002, pp. 117–128.

[5] N. Beckmann, H-P. Kriegel, R. Schneider, and B. Seeger, "The r*-tree: an efficient and robust access method for points and rectangles," in *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, 1990, pp. 322–331.

[6] S. Berchtold, C. Böhm, and H. P. Kriegel, "The pyramid-tree: breaking the curse of dimensionality," in *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, 1998, pp. 142–153.

[7] R. Weber, H. J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *Proc. of Int. Conf. on Very Large Data Bases*, 1998, pp. 194–205.

[8] P. Ciaccia and M. Patella, "Pac nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces," in *Proc. of Int. Conf. on Data Engineering*, 2000, pp. 244–255.

[9] C. Li, E. Chang, M. Garcia-Molina, and G. Wiederhold, "Clustering for approximate similarity search in high-dimensional spaces," *IEEE Trans. on Knowledge and Data Engineering*, vol. 14, no. 4, pp. 792–808, 2002.

[10] K. P. Bennett, U. Fayyad, and D. Geiger, "Density-based indexing for approximate nearest-neighbor queries," in *Proc. of Conf. on Knowledge Discovery in Data*, 1999, pp. 233–243.

[11] S-A. Berrani, L. Amsaleg, and P. Gros, "Approximate searches: k-neighbors + precision," in *Proc. of Int. Conf. on Information and knowledge management*, 2003, pp. 24–31.

[12] P. Ciaccia and M. Patella, "Approximate similarity queries: A survey," Tech. report, University of Bologna: MultiMedia DataBase Group, 2001.

[13] P. Ciaccia and M. Patella, "Pac nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces," in *Proc. of Int. Conf. on Data Engineering*, 2000, pp. 244–255.

[14] A. Joly, C. Frélicot, and O. Buisson, "Feature statistical retrieval applied to content-based copy identification," in *Int. Conf. on Image Processing*, 2004.

[15] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.

[16] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *Proc. of Int. Conf. on Computer Vision*, 1993, pp. 231–236.