Statistical similarity search applied to content-based video copy detection

Alexis Joly[(1)], Olivier Buisson[(2)], Carl Frélicot[(3)]

[(1)] Projet IMEDIA, INRIA Rocquencourt - B.P. 105 78153 Le Chesnay Cedex, France
[(2)] Département Recherche et Études, Institut National de l'Audiovisuel, 94366 Bry/Marne, France
[(3)] Lab. d'Informatique – Image – Interaction, Université de La Rochelle, 17042 La Rochelle, France

Abstract— **Content-based copy detection (CBCD) is one of the emerging multimedia applications for which there is a need of a concerted effort from the database community and the computer vision community. Recent methods based on interest points and local fingerprints have been proposed to perform robust CBCD of images and video. They include two steps: the search of similar fingerprints in the database 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 of similar features in the database 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 introduce a new approximate search paradigm, called *Statistical Similarity Search* (S^3), dedicated to local fingerprints and we describe the original indexing structure we have developped to compute efficiently the corresponding queries. The key-point relates to the distribution of the relevant fingerprints around a query. Since a video query can result from (a combination of) more or less transformations of an original one, we modelize the distribution of the distorsion vector between a referenced fingerprint and a candidate one. Experimental results show that these statistical queries allow high performance gains compared to classical ϵ-range queries. By studying the influence of this approximate search on a complete CBCD scheme based on local video fingerprints, we also show that trading quality for time during the search does not degrade seriously the global robustness of the system, even with very large databases including more than 20,000 hours of video.

I. 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]. As opposed to watermarking, the CBCD approach only uses a content-based comparison between the original and the candidate objects. For storage and computational considerations, it generally consists in extracting as few features as possible from the candidate objects and matching them with a database (DB). Since the features must be discriminant enough to identify an image or a video, the features are often called *fingerprints* or *signatures* [4]. CBCD presents two major advantages. First, a video clip which has already been delivered can be recognized. Secondly, content-based features are intrinsically more robust than inserted watermarks because they contain information that cannot be suppressed without considerably corrupting the perceptual content itself.
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 of similar local fingerprints in the database 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 (see section III). 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 cost of the similarity search in the fingerprints reference database, which can be very large. In its essence, the *similarity query* paradigm is to find objects in the database which are similar to a given query object up to a given degree [5]. In order to assess the sim-

ilarity between two objects, a distance is generally used to perform *k-nearest neighbors queries* or ϵ-*range queries*.

To solve this problem, most multimedia retrieval systems use multidimensional index structures such as R-tree family techniques ([6], [7], [8]). To overcome the *dimensionality curse* phenomenom that occurs when the dimension of the descriptors becomes very high, other index structures have been proposed, e.g. the pyramid tree [9] or dimension reduction techniques [10]. Sometimes, improved sequential techniques such as the VA-file are even more profitable than all other structures [11]. However, the search time remains to high for many of the emerging multimedia applications [12]. For the last few years, researchers are interested in trading quality for time [13], [14], [15], [16], [17] and the paradigm of *approximate similarity search* has emerged [18]. Some of the proposed solutions are simply *early stopping approaches* [15], [14]. The search is stopped when a fixed number of the most relevant bounding regions have been visited. The precision of the search is therefore not controlled. Other techniques, e.g. [19], are based on geometrical approximations during the filtering rules of the search algorithm. They guarantee a priori the maximum the distance between the approximate results and the exact k nearest neighbors. More recent techniques are based on a probabilistic selection of the bounding regions used in the indexing structure [16], [17]. They allow to control directly the expected percentage of the real k-nearest neighbors.

The previous approximate search techniques are dedicated only to k-nearest neighbor queries. 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 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 database, 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 propose is to extend the approximate search paradigm to ϵ-range queries, leading to the *statistical query* paradigm. It is introduced in section II. In section III, we describe the content-based video copy detection scheme based on local fingerprints. The index structure we have developped to compute the statistical queries is presented in section IV and experimental results showing its efficiency are provided in section V. Concluding remarks and some perspectives of this work are given in section VI.

II. Statistical query

By excluding several regions of an hyper-spherical query, having a too small intersection with the bounding regions of the index structure, it is indeed 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 hyper-sphere but this is not true in reality, as illustrated on figure 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 $w_{scale} = 0.8$, 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. The figure shows that 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:

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

where $S\left(m\right)$ is the fingerprint of a referenced pattern m, $S\left(t(m)\right)$ 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 of it. We defined the *statistical query of expectation* α as the search of all

the fingerprints contained in a region V_α of the feature space satisfying:

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

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

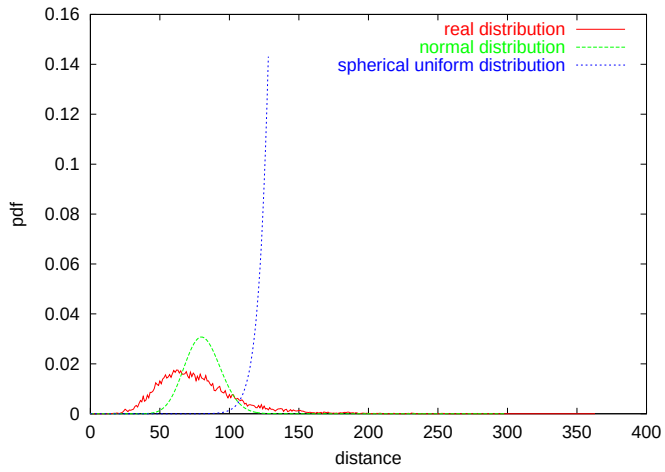


Fig. 1. Distribution of the distance between a fingerprint and its distorted version after resizing of a video sequence ($w_{scale} = 0.8$)

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.

III. Video local fingerprints and voting strategy

In this section, we briefly remind the main steps of the video CBCD scheme proposed in [3]. In order to be robust to inserting and shifting, which are frequent operations in the TV context, the method is based on local descriptors as suggested in [20] or [21]. The local fingerprints extraction includes the three following steps:

- a key-frame detection, based on the mean of the frames difference also called *intensity of motion*. A gaussian filter is applied to it and the key-frames are selected at the extrema positions of the resulting signal.
- an interest point detection in each key-frame, processed by an improved version of the Harris detector [21].
- a local characterisation computed around each interest point, leading to the following $D = 20$-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:

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

The fingerprints are coded with one byte for each component. Thus, all measures in the rest of the paper refer to the space $[0, 255]^{D=20}$.

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 $id \in \{Id_{jk}\}_{k \in K_j;\, j \in [1, N_{cand}]}$, 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 [22] for more 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.

IV. Indexing Structure

The structure we use to index the fingerprints and to perform the statistical queries is based on a Hilbert's space filling curve. Multidimensional indexing using Hilbert's space filling curves was originally suggested by Faloutsos [23] and fully developed by Lawder [24]. The main advantage of such index structures is to convert the complex multidimensional access problem to a simple 1-dimensional access problem. The principle of our index structure is quite similar to Lawder's approach: first, the query is mapped to Hilbert's curve coordinate and is converted into several curve sections according the filtering rules of the search. Then, a refinement step consists in scanning the fingerprints belonging to these sections. The Hilbert's curve clustering property limits the number and the dispersion of these sections reducing the number of memory accesses.

However the method we propose differs in several main points. Lawder's filtering step requires the use of state diagrams to compute the mapping to the Hilbert's curve which limits the dimension to about 10 because of primary storage considerations. The proposed method uses the so-called Butz algorithm [25] for the mapping and requires little memory. Furthermore, only hyper-rectangular range queries are computable with Lawder's indexing technique. Spherical range queries (ϵ-range queries) or statistical queries are possible with our new structure.

The fingerprints database is physically ordered according to the position of the fingerprints on the Hilbert's curve. That implies that the S^3 system is static: no dynamic insertion or deletion are possible. For a given query, once the curve sections have been identified by the statistical filtering rules (see subsection IV-A), the corresponding sets of successive fingerprints in the database are localized by an simple index table. Then, the refinement step sequentially scans each set of successive fingerprints like a classical sequential scan. The fingerprint database is stored in a single file but it is entirely loaded in primary storage at the start of the detection stage of our CBCD system. When the DB exceeds primary storage size, it is cyclically loaded in several memory size blocks and several queries are searched together (see subsection IV-B). For computational efficiency, the main assumption of the S^3 system is that the D components of the distorsion vector are independent:

$$p_{\Delta S} = \prod_{j=1}^{D} p_{\Delta S j}$$

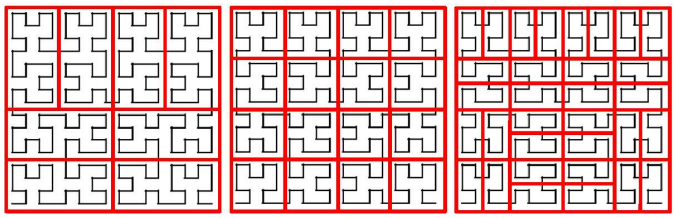


Fig. 2. Space partition induced by the Hilbert's space filling curve for $D = 2$ and $K = 4$ at different depths – from left to right: p=3,4 and 5

A. Statistical filtering step

The K-th order approximation of Hilbert space-filling curve in a D-dimensional grid space H_K^D is a bijective mapping: $\left[0, 2^K - 1\right]^D \leftrightarrow \left[0, 2^{KD} - 1\right]$. The main property is that two neighboring intervals on the curve always remain neighboring cells in the grid space. The reciprocal property is generally not true and the quality of a space filling curve can be evaluated by its ability to preserve a certain locality on the curve. Some intermediate variables of the Butz algorithm allow to easily define the space partition corresponding to the regular partition of the curve in 2^p intervals [26]. Parameter $p \in [1, KD]$ is called the *depth* of the partition by analogy to KD-trees. As illustrated on Figure 2, the space partition is a set of 2^p hyper-rectangular blocks (called p-blocks) of same volume and shape but of different orientations.

For a p-partitioned space and a candidate fingerprint Q, the statistical query inequality (1), may be satisfied by finding a set B_α of p-blocks such as:

$$\sum_{i=1}^{card(B_\alpha)} \int_{b_i} p(X - Q)\, dX \geq \alpha \qquad (3)$$

where $B_\alpha = \{b_i : i \in [1, card(B_\alpha)]\}$ and $0 \leq$

$card(B_\alpha) \leq 2^p$.

In practice, $card(B_\alpha)$ should be minimum to limit the cost of the search. We refer to this particular solution as B_α^{min}. Its computation is not trivial because sorting the 2^p blocks according to their probability is not affordable. Nevertheless, it is possible to identify quickly the minimal set of blocks with a total probability greater than a fixed threshold t:

$$B(t) = \left\{ \{b_i\} : \int_{b_i} p(X - Q)\, dX > t \right\}$$

and the corresponding probability sum:

$$P_{sup}(t) = \sum_{i=1}^{card(B(t))} \int_{b_i} p(X - Q)\, dX$$

Since $card(B(t))$ decreases with t, finding B_α^{min} is equivalent to finding t_{max} verifying:

$$\begin{cases} P_{sup}(t_{max}) \geq \alpha \\ \forall t > t_{max}, P_{sup}(t_{max}) < \alpha \end{cases} \quad (4)$$

As $P_{sup}(t)$ also decreases with t, t_{max} can be easily approximated by a method inspired by Newton-Raphson technique.

Parameter p is of major importance since it directly influences the response time of our approximate method

$$T(p) = T_f(p) + T_r(p)$$

The filtering time $T_f(p)$ is strictly increasing because the computation time of B_α and the number of memory accesses increase with p. The refinement time $T_r(p)$ is decreasing because the *selectivity* of the filtering step increases, i.e $card(S_\alpha)$ decreases with p. The response time $T(p)$ has generally only one minimum at p_{min} which can be learned at the start of the retrieval stage in order to obtain the best average response time.

B. Pseudo-disk strategy

When the fingerprints database exceeds memory size, a fixed number N_{sig} of fingerprints are searched together. At the start of the retrieval stage, the Hilbert's curve is split in 2^r regular sections ($0 \leq r \leq p$), such that the most filled section fits in memory. The filtering step, which is independent of the database, is processed for each fingerprint during a first stage. Each of the 2^r sections is then sequentially loaded in main memory and the refinement step is processed for the N_{sig} fingerprints. The average total response time per query is given by:

$$\overline{T}_{tot} = \overline{T} + (T_{load}/N_{sig}) \quad (5)$$

where T_{load} is the loading time for the entire DB. This additional time introduces a linear component in the response time against DB size. However it can be neglected in most cases by adjusting N_{sig}. In our CBCD system, N_{sig} is automatically set to obtain an average loading time that is sublinear with the database size.

C. Distortion model and assessment

The only necessary assumption for the distortion probabilistic model is the independence of the components. However, we use in practice a zero-mean normal distribution with the same standard deviation σ whatever the component:

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

For a given image transformation, the single parameter to be set σ can be estimated on a set of video sequences by simulating a perfect interest points detector, the points position in the transformed sequence being computed according the position in the original sequence. Let σ_j be the standard deviation of the j^{th} component of the distorsion vector. We use the mean $\overline{\sigma}$ of the D standard deviation estimates $\widehat{\sigma}_j$ as an estimate for σ.

The relevance of this model is simply tested by the observed retrieval rate R obtained using the S^3 technique for different values of the query expectation α. The transformation applied to the video sequences was a combinaison of several transformations: resizing, gamma modification, noise addition and a simulated imprecision in the position of the interest points by shifting the theorical position by 1 pixel. The results are shown in Figure 3. Since the error does not exceed 7%, we validate the model. A more sophisticated model should certainly improve this precision but we remind that we only aim at showing that statistical queries are an alternative to a query geometrical approximation which leads to an unreallistic model. Furthermore, we will see in the experiments section that the main asset of the statistical queries is that the lost of quality during the search does not seriously degrade the global robustness of the CBCD system. Thus, a fine control of the precision of the search is not an essential objective.

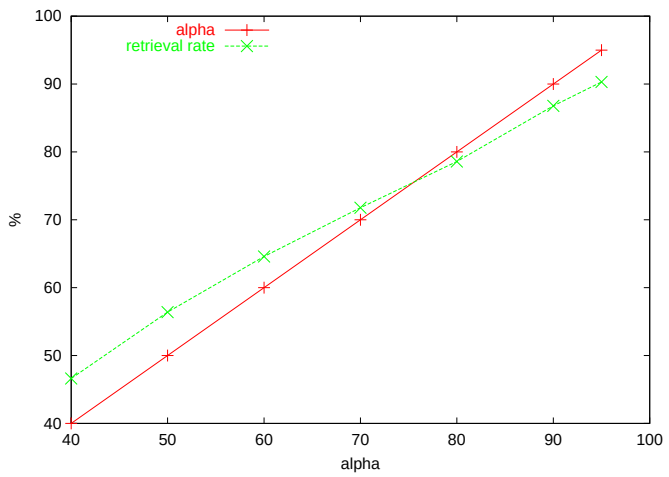


Fig. 3. Retrieval rate R obtained using the S^3 technique *vs*. expectation α of the statistical query

transformation	$\overline{\sigma}$	R
$w_{scale} = 0.84$, $\delta_{pix} = 1$	**23.43**	**80.74**
$w_{scale} = 1.26$, $\delta_{pix} = 1$	21.95	81.69
$w_{scale} = 0.91$, $\delta_{pix} = 1$	18.13	92.49
$w_{scale} = 0.98$, $\delta_{pix} = 1$	14.92	97.30
$w_{gamma} = 2.08$, $\delta_{pix} = 1$	12.17	98.52
$w_{gamma} = 0.82$, $\delta_{pix} = 1$	9.23	99.79
$w_{noise} = 10.0$, $\delta_{pix} = 0$	6.60	99.65

TABLE I
DETECTION RATE R FOR TRANSFORMATIONS OF DECREASING SEVERITY - $\alpha = 85\%$ AND $\sigma = 23.43$

In reality, the distribution of the relevant similar fingerprints depends on the transformations distribution. However, the appearance frequency of a transformation is generally not known and taking this distribution into account, would be equivalent to exclude the most severe transformations. Thus, we consider that the distortion model only concerns the distribution obtained for the most severe transformation. A statistical query of expectation α for the most severe transformation should guarantee a better expectation for all other transformations.
We propose to take the value of $\overline{\sigma}$ as a *severity criterion*. Table I reports some retrieval rate R obtained for different transformations with decreasing severities $\overline{\sigma}$. The expectation α is 85% and the reference severity is the one corresponding to the most severe transformation, i.e. $\overline{\sigma} = 23.43$. The detection rate for this reference transformation is 80.74% and we verify that it is always higher than this value for the other transformations. Except for last transformation, the retrieval rate R increases when the severity of the transformation decreases. Note that, in practice, it is not necessary to know exactly the most severe transformation. Indeed, the value of σ allows to make a compromise between the robustness of the CBCD and the search time. However, there is a limit for which it becomes useless to increase σ since the interest point detector repeatability will be close to zero for transformations that are too severe.

V. EXPERIMENTS AND RESULTS

Experiments were computed on a Pentium IV (CPU 2.5 GHz, cache size 512Kb, RAM 1.5 Gb). Response times were obtained with unix `getrusage()` command. The five kinds of transformations studied in these experiments are illustrated in Figure 4.

All the databases contain real video local fingerprints extracted by the method described in section III. The referenced video sequences come from the so called *SNC* database stored at the French *Institut National de l'Audiovisuel* (INA), whose main missions include collecting and exploiting French television programs (200,000 hours of digitized television archives). The *SNC* video sequences are stored in $MPEG1$ format with an image size of 352×288. They contain all kinds of TV programs from the Fourties at our days: news, sport, show, variety, films, reports, black&white archives, advertissements, ... They unfortunately also contain noise, black sequences, test cards which degrade a little some of our experimental evaluations. During more than one year, we have computed the fingerprints of randomly selected video sequences from the *SNC* database resulting in about $75,000$ hours of video fingerprints. The average number of local fingerprints per hour of video is about $50,000$. Thus, a database 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).

A. 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 database and construct 1000 queries $Q = S + \Delta S$, where the components of the distorsion vector ΔS are

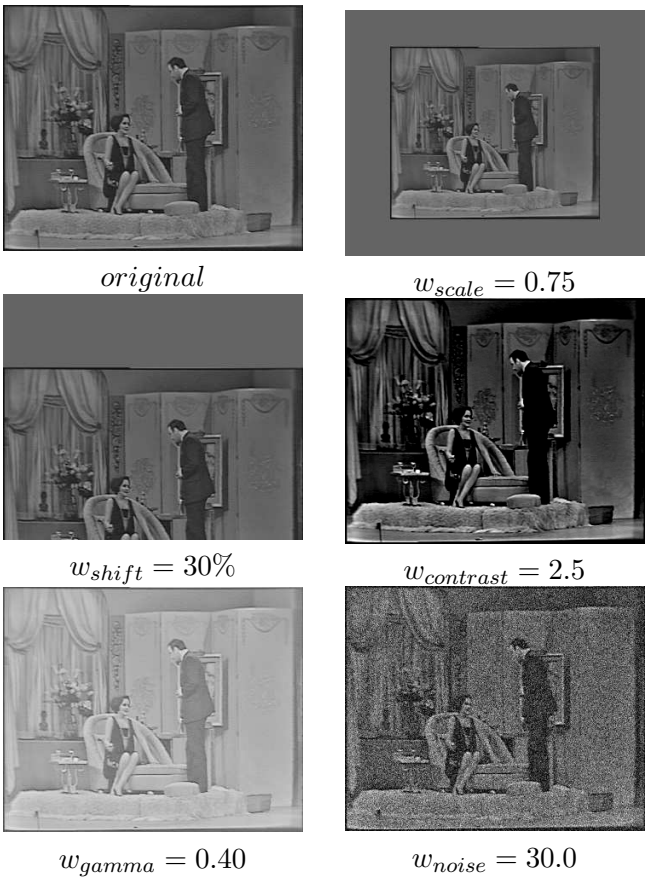


Fig. 4. The five kinds of transformations studied in the experiments: resize, shift, contrast, gamma, noise addition

independently generated according a zero mean normal distribution $p_{\Delta Sj}(x_j) = f_{\mathcal{N}(0,\sigma_Q)}(x_j)$ with $\sigma_Q = 18.0$. These queries are then searched in the database using the proposed index structure with 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 (Figure 6) and the retrieval rate (Figure 5), 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:

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

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 Figure 6 (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. This is due to the less number of p-blocks intercepted by the statistical query. The geometrical constraint of an exact ϵ-range query degrades seriously the search time without improving the retrieval rate, as shown in Figure 5. 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.

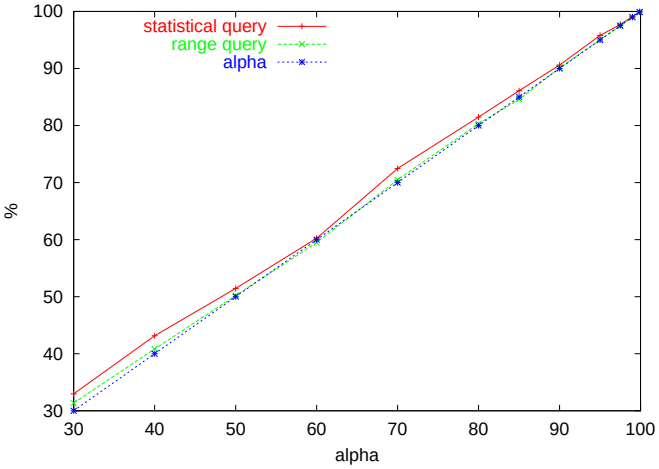


Fig. 5. Retrieval rate (%) *vs.* query expectation α

B. Behavior when faced to a very large database

We aim at showing the performance of the proposed scheme when the database size increases. The search time is compared to those of a sequential scan, which is a reference method. We implemented our own version of the sequential scan so that the two methods are comparable. About 200,000 candidate fingerprints in $[0,255]^{20}$ are extracted from a french television video stream and searched in the DB using the S^3 technique. Only 1000 of these fingerprints are searched using the sequential scan method because it is too much time-consuming. The parameters for the statistical query approach are $\alpha = 80\%$ and $\sigma = 20.0$. For

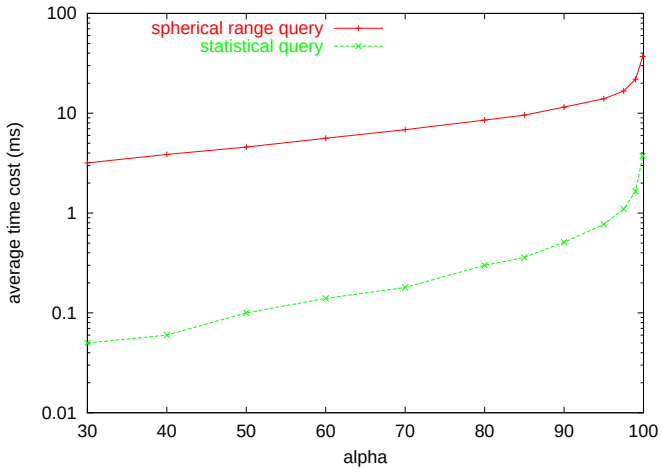


Fig. 6. Average search time(ms) *vs.* query expectation α

the sequential scan, the range query parameter ϵ is set to 93.6 so that both search methods are comparable (same expectation). Thirteen DB of exponentially growing sizes are used, the smallest containing $77,131$ fingerprints (about 1 hour of video) and the largest containing $1,543,902,419$ fingerprints (about $30,000$ hours of video). The obtained average search times are shown in Figure 7. Both the DB sizes and times are displayed in logarithmic coordinates, thus the linear relationship between the average search time of the sequential scan and the DB size still remains.

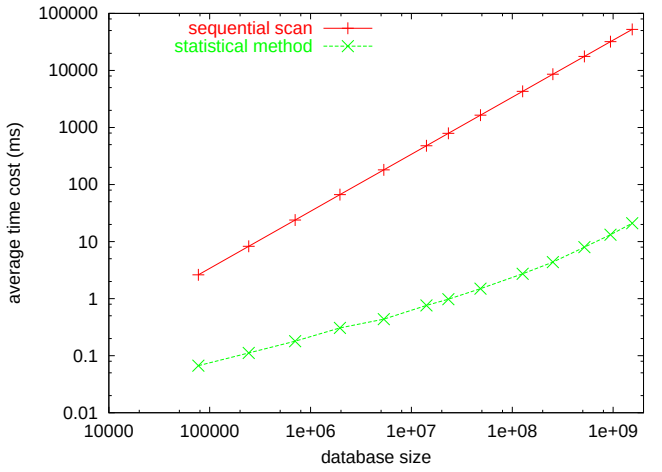


Fig. 7. Average search time (ms) *vs.* database size

As expected, the proposed approach outperforms the sequential one. One can observe that when the DB size do not exceed 50 millions fingerprints, the average search time of the S^3 method is a sub-linear functions of the DB size with a constant slope in logarithmic coordinates. This means that the gain against the sequential scan method increases exponentially. For larger databases, the slope of the S^3 curve in logarithmic coordinates is increasing and reaches the slope of the sequential scan method. This is because of the main memory storage in the former case and the pseudo-disk strategy in the latter case resulting to an additional linear component. Thus, the obtained gain tends to a constant. For the largest DB, the proposed scheme is more than $2,500$ times faster.

C. Robustness of the video CBCD system

The purpose of these last 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 databases and apply to them the five kinds of transformations presented in Figure 4:

- resize of factor w_{scale}
- vertical shift of w_{shift} % of the image
- gamma modification: $I'(x,y) = I(x,y)^{w_{gamma}}$
- contrast modification: $I'(x,y) = w_{contrast} I(x,y)$
- gaussian noise addition with standard deviation w_{noise}

The 100 transformed sequences are submitted as candidates to the video CBCD system and we measure the good detection rate, a candidate video sequence being well detected if at least one of its key images is well identified with a tolerance of 2 frames and if the similarity measure n_{sim} is higher than a decision threshold. This threshold is 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 statistical query 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 in Figure 8 and 9 in the form of abacuses of:

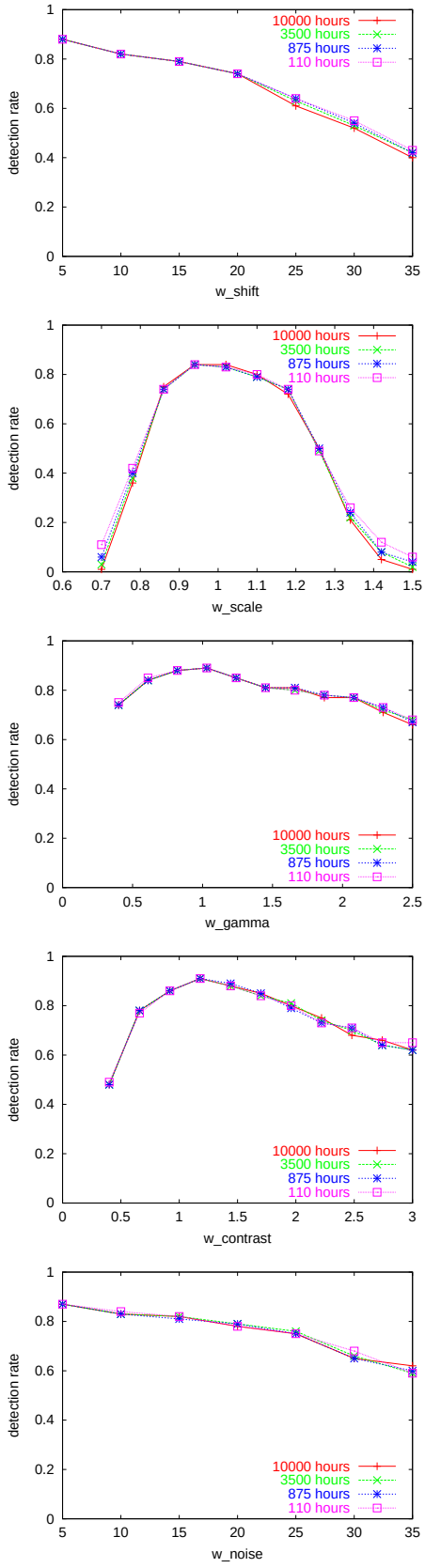
- the DB size (Figure 8) with α set to 80%
- α (Figure 9) using a DB containing about 3500 hours of video

for the five transformations. 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 each figure.

Before discussing the obtained results, we bring the following precision in connection with the detection rates. Even when the transformation is light, the detection rate does not exceed 90% mainly because of the length of the candidate sequences (10 seconds). They can reach 100% when extended to 30 seconds

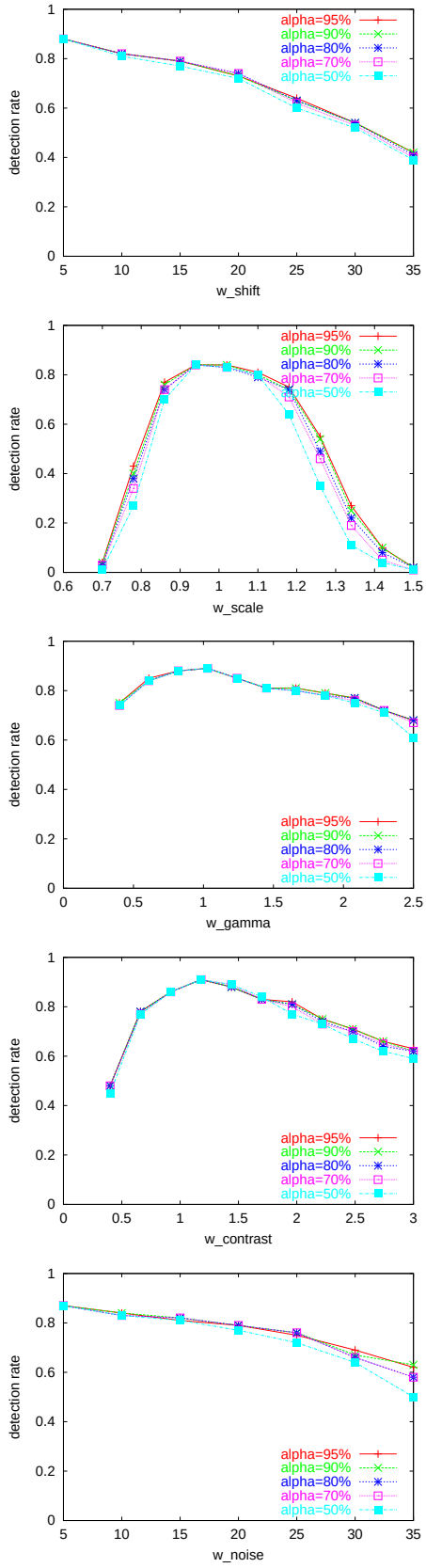
length. Another reason is the databases themselves that contain about 2% of irrelevant sequences like black or noisy sequences whose discrimination is very hard.

Subplots of Figure 8 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.
Subplots of Figure 9 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 transformations, 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.



video hours	fingerprints number	search time (ms)
110	6,013,876	0.69
875	49,107,362	1.81
3500	195,103,572	3.45
10,000	555,238,372	9.11

Fig.8. DB size abacuses



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

Fig.9. α abacuses

D. TV monitoring system

For several months, a video CBCD system based on the S^3 technique and local fingerprints approach described in this paper is continuously monitoring a french TV channel with a reference DB including more than 20,000 hours of archives. The average monitoring time is 2 times faster than real time. Some examples of detections illustrating that the CBCD system is robust to studied transformation are shown in Figure 10.

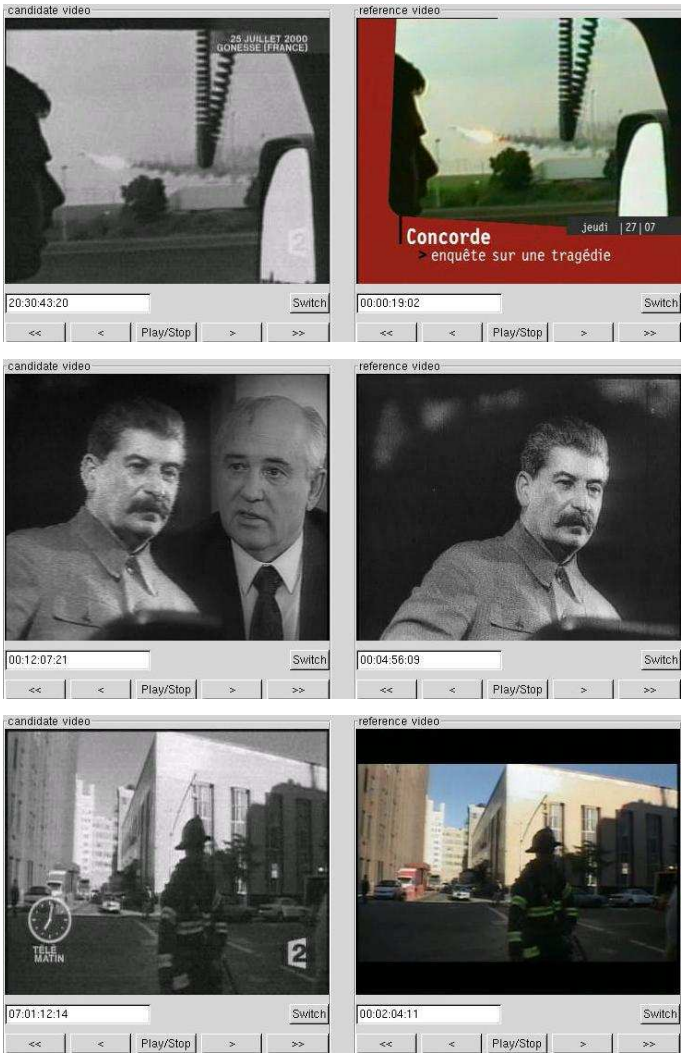


Fig. 10. CBCD examples: TV candidate sequences captured in black and white (left) and identified sequences in the DB (right)

VI. Conclusion and Perspectives

In this paper, we have proposed a new approximate search paradigm based on statistical queries and we have introduced the indexing structure to compute them efficiently. The resulting technique, called *Statistical Similarity Search* (S^3) technique, is based on

the distorsion vector modelization of the fingerprint. Even if such distorsion vectors are more difficult to interprete in the general framework of content-based retrieval, they allow to take into account the more or less strong transformations that a video query can have undergone. It is therefore particularly adapted to content-based copy detection.

We show that the geometrical constraint of a classical exact ϵ-range query can degrade seriously the search time without systematically increase the retrieval rates. The use of a statistical filtering of the bounding regions instead of an exact geometrical filtering can lead to a 100 times faster search with comparable retrieval rates. The improvement is even more important in case of large databases, for which we observed that the S^3 technique can be $2,500$ times faster than a sequential scan. We studied the influence of the proposed search strategy on our *Content-Based Copy Detection* ($CBCD$) system based on local fingerprints retrieval and a voting strategy. The system is rather robust and allows us to conclude that trading quality for time during the search is highly profitable, even when the size of the database becomes very large.

Investigations in the statistical modeling of the distortion vector, including the component independence assumption, should probably improve the efficiency and the precision of the proposed statistical similarity search. However, the main future works will concern the voting strategy for two reasons. Firstly, when the databases becomes very large, the number of retrieved fingerprints by the CBCD system during the search increases seriously and this step will probably become a new bottleneck of the total search time. Secondly, we would like to extend the estimation step to the spatial positions of the interest points in order to improve the discriminance of the fingerprints.

Notice: the technique described in this paper is being patented by INA, which has all commercial rights.

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, C. Frélicot, and O. Buisson, "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] C. Böhm, S. Berchtold, and D. A. Keim, "Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases," *ACM Computing surveys*, vol. 33, no. 3, pp. 322–373, 2001.

[6] 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.

[7] S. Berchtold, D. A. Keim, and H-P. Kriegel, "The x-tree: An index structure for high-dimensional data," in *Proc. of Int. Conf. on Very Large Data Bases*, 1996, pp. 28–39.

[8] N. Katayama and S. Satoh, "The sr-tree: An index structure for high-dimensional nearest neighbor queries," in *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, 1997, pp. 369–380.

[9] 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.

[10] C. Faloutsos and K-I. Lin, "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," in *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, 1995, pp. 163–174.

[11] R. Weber and S. Blott, "An approximation based data structure for similarity search," Tech. report, ESPRIT project HERMES (no. 9141), 1997.

[12] L. Amsaleg, P. Gros, and S-A. Berrani, "Robust object recognition in images and the related database problems," *Special issue of the Journal of Multimedia Tools and Applications*, vol. 23, pp. 221–235, 2003.

[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] R. Weber and K. Böhm, "Trading quality for time with nearest neighbor search," in *Proc. of Int. Conf. on Extending Database Technology*, 2000, pp. 21–35.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] L. Amsaleg, P. Gros, and S-A. Berrani, "A robust technique to recognize objects in images, and the db problems

it raises," in *Proc. of Int. Workshop on Multimedia Information Systems*, 2001, pp. 1–10.

[21] 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.

[22] 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.

[23] C. Faloutsos and S. Roseman, "Fractals for secondary key retrieval," in *Proc. of ACM Symp. on Principles of database systems*, 1989, pp. 247–252.

[24] J. K. Lawder and P. J. H. King, "Querying multi-dimensional data indexed using the hilbert space-filling surve," *SIGMOD Record*, vol. 30, no. 1, pp. 19–24, 2001.

[25] A. R. Butz, "Alternative algorithm for hilbert's space-filling curve," *IEEE Trans. on Computers*, vol. C, no. 2, pp. 424–426, 1971.

[26] J. Lawder, "The application of space-filling curves to the storage and retrieval of multi-dimensional data," Phd thesis, University of London, 1999.