Discriminant Local Features Selection using Efficient Density Estimation in a Large Database

Alexis Joly
INRIA-IMEDIA
78153 Le Chesnay
France
alexis.joly@inria.fr

Olivier Buisson
INA
94366 Bry-sur-Marne
France
obuisson@ina.fr

ABSTRACT

In this paper, we propose a density-based method to select discriminant local features in images or videos. We first introduce a new fast density estimation technique using a simple grid index structure and specific queries based on the energy of the gaussian function. This method enables the nonparametric density estimation of target features with very large sets of source features. We then apply it to the selection of discriminant local features: the principle is to keep only the features having the lowest density in a feature database constructed from a large collection of representative objects (images or videos). Experiments are reported to evaluate the density estimation technique in terms of both quality and speed. The density-based selection of discriminant local features is evaluated in a complete video content-based copy detection framework using Harris interest points.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Miscellaneous

General Terms

Algorithms,Performance,Experimentation

1. INTRODUCTION

In our previous work [8], we proposed an efficient similarity search technique to retrieve multidimensional *distorted* features in very large databases. Instead of using a usual search paradigm such as range query or K-nearest neighbors query, we introduced a new approximate search paradigm using distortion-based probabilistic queries. Given a query $\boldsymbol{X}$, the principle of the search is to retrieve all the features of the database contained in a region V_α of the feature space satisfying:

$$\int_{V_\alpha} p_{\boldsymbol{\Delta S}}\left(\boldsymbol{Z}-\boldsymbol{X}\right) \boldsymbol{dZ} \geq \alpha \qquad (1)$$

where $p_{\boldsymbol{\Delta S}}(.)$ is the probability density function of the *distortion* $\boldsymbol{\Delta S} = \boldsymbol{S} - \boldsymbol{S_d}$, i.e the difference vector between an original feature S and a distorted feature S_d (typically obtained after transformation of an image).
In this paper, we propose to apply this technique to approximate density estimation in large databases. By replacing the probability density function $p_{\boldsymbol{\Delta S}}$ by a gaussian function, our method is indeed well suited to select only the features of the database having a significant contribution to a density estimation function based on a gaussian kernel.
The other and main contribution of this paper consists in a new interest points selection strategy. Instead of maximizing an operator using only the local information content, the method selects the points according to the density of their local features in a representative source database. The idea is to detect *rare* local features that should be more discriminant and thus useful in an indexing perspective. Our proposed nonparametric density estimation technique being very fast, it enables the density estimation in very large source datasets even in a real-time video context as discussed in the experiments.
The paper is organized as follows: Section 2 discusses the issues of fast nonparametric density estimation and describes our method. Section 3 presents its application to the selection of relevant interest points in images and video clips. Experiments are reported in section 4.

2. FAST NON-PARAMETRIC DENSITY ESTIMATION

2.1 Issues

Nonparametric density estimation is an attractive technique in many computer vision applications since it does not require any assumption on the forms of the underlying probability density function. Furthermore, practical computer vision problems often involve high dimensional multimodal densities which can not be simply represented. A nonparametric technique which is quite general is the kernel density estimation technique [17]. In this technique the underlying probability density function is estimated as:

$$\widehat{f}(\boldsymbol{X}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{X} - \boldsymbol{X_i})$$

MIR'05, November 10–11, 2005, Singapore.

where K is a kernel function (typically a Gaussian) centered at the data points X_i $(i = 1, ..., N)$ and α_i are weighting coefficients (typically uniform weights are used, i.e., $\alpha_i = 1/N$). The use of such an approach requires a way to efficiently evaluate the estimate $\widehat{f}(\boldsymbol{X})$ at any new point X. In general, given N original data samples (*sources*) and M points at which the density must be evaluated (*targets*), the complexity is $O(NM)$ evaluations of the kernel function, multiplications and additions. For many applications in computer vision, where both real-time operation and generality of the classifier are desired, this complexity can be a significant barrier to the use of these density estimation techniques. Several methods, however, allow to speed up the process. When both the number of targets and the number of sources are high, the fast Gauss transform [4] and the dual tree [3] are known to be the most efficient techniques. In this paper, we are more interested in estimating very quickly the density of one single target feature faced with a very large source dataset in order to deal with in line image retrieval or real-time video context. In this case, the most commonly used techniques are range query searching algorithms performed in indexing structures such as grids, classic kd-trees [3] or also an Anchor hierarchy [14].
When using a gaussian kernel and uniform weights, the density function $\widehat{f}(\boldsymbol{X})$ can be simply expressed as:

$$\widehat{f}(\boldsymbol{X}) = \frac{1}{N} \frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \sum_{i=1}^{N} e^{\frac{||\boldsymbol{X}-\boldsymbol{X_i}||^2}{2\sigma^2}} \tag{2}$$

where σ is the bandwidth of the estimation and D the dimension of the feature space. The technique proposed in this paper is only dedicated to this simplified density estimation problem although it could be easily extended to non uniform weights.

2.2 Proposed technique

2.2.1 Principle

Given a target vector $\boldsymbol{X}$ where the density has to be estimated, the principle of our technique is to consider only the N_α features $\boldsymbol{X_i}$ of the source database contained in a region $V_\alpha(\boldsymbol{X})$ of the feature space satisfying:

$$\frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \int_{V_\alpha(\boldsymbol{X})} e^{\frac{||\boldsymbol{X}-\boldsymbol{Z}||^2}{2\sigma^2}} \boldsymbol{dZ} \geq \alpha \tag{3}$$

We refer to this search paradigm as an α-query. The density estimation function $\widehat{f}_\alpha(\boldsymbol{X})$ is then computed only on the N_α selected source features (i.e the features $\boldsymbol{X_i}$ belonging to $V_\alpha(\boldsymbol{X})$):

$$\widehat{f}_\alpha(\boldsymbol{X}) = \frac{1}{N} \frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \sum_{i=1}^{N_\alpha} e^{\frac{||\boldsymbol{X}-\boldsymbol{X_i}||^2}{2\sigma^2}} \tag{4}$$

Intuitively, our technique guaranties that α percent of the energy of the gaussian function will be included in the estimation. In practice, V_α is a set of bounding regions depending on the space-partition on which the index structure is based (see section 2.2.2). The main difference with usual fast estimation techniques based on KD-trees or grid partitions [3], is that the pruning of the bounding regions is not based on distance computations. In these methods, a bounding box is pruned from the priority queue if the kernel value between the closest point of the bounding region and the target vector is under a certain threshold. The used algorithms are mainly range query searching algorithms [3]. In our method, the pruning of the data chunks is not based on geometrical rules but on a global criterion making equal to α the percentage of the energy of the gaussian function recovered by the selected bounding regions. The underlying heuristic is that most of the features having a significant contribution to the density estimation function (Equ. 2) belong to the selected regions. As V_α is determined such as it minimizes the number of bounding regions, this approach allows to have significantly less data chunks to process. We showed in [9], that the number of visited data chunks in a 20 dimensional grid partition was 100 times lower when using an α-query instead of a range query, both recovering the same percentage of the energy of the gaussian function (see Fig. 1 for an illustration).

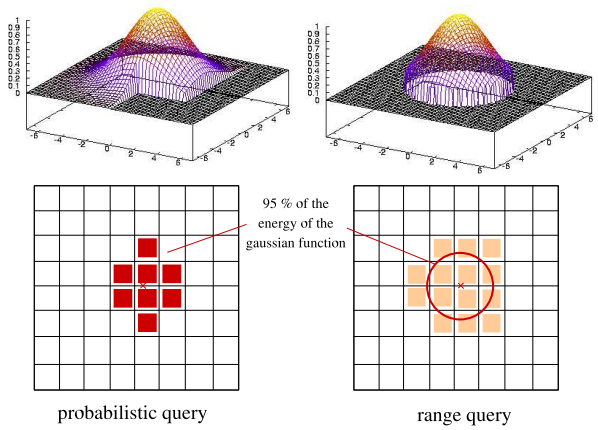

Figure 1: Difference between a probabilistic query and a range query

2.2.2 Indexing structure and density estimation algorithm

The indexing structure we use to process our α-queries is described in [8]. The space-partition is a simple grid induced by the regular split of a Hilbert space-filling curve as illustrated on Figure 2. It results in a set of 2^p non overlapping and hyper-rectangular bounding regions, called *p-blocks*, which are well-suited to compute quickly the integral of the gaussian function. The depth p of the partition is equal to the number of bits of the Hilbert derived keys used to access the data pages corresponding to each block. The density estimation algorithm is composed of two steps: a filtering step that selects the relevant p-blocks and a refinement step that exhaustively processes all the features belonging to the selected blocks. Thanks to the separability property, the integral of the gaussian function on a p-block b can be easily computed as:

$$\frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \int_b e^{\frac{||\boldsymbol{X}-\boldsymbol{Z}||^2}{2\sigma^2}} \boldsymbol{dZ} = \frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \prod_{j=1}^{D} \int_{u_j^i}^{v_j^i} e^{\frac{(\boldsymbol{x_j}-\boldsymbol{z_j})^2}{2\sigma^2}} dz_j$$

where u_j^i and v_j^i are the lower and upper bounds of the p-block b along the j^{th} axis, x_j and z_j are the j^{th} component of respectively the target feature $\boldsymbol{X}$ and any vector $\boldsymbol{Z}$.

For a p-depth partitioned space and a target feature $\boldsymbol{X}$,

inequality (3), may be satisfied by finding a set B_α of p-blocks such as:

$$\frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \sum_{i=1}^{card(B_\alpha)} \int_{b^i} e^{\frac{||\boldsymbol{X}-\boldsymbol{Z}||^2}{2\sigma^2}} \boldsymbol{dZ} \geq \alpha \qquad b^i \subset B_\alpha, \forall i \quad (5)$$

where $card(B_\alpha) \leq 2^p$ is the number of blocks in B_α. 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 weights is not affordable. Nevertheless, it is possible to quickly identify the set $B(\tau)$ containing all the blocks for which the integral of the gaussian function is greater than a fixed threshold τ:

$$B(\tau) = \left\{ \left\{ b^i \right\} / \frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \int_{b^i} e^{\frac{||\boldsymbol{X}-\boldsymbol{Z}||^2}{2\sigma^2}} \boldsymbol{dZ} > \tau \right\}$$

The total energy of the gaussian function contained in $B(\tau)$ is given by:

$$P_\Sigma(\tau) = \frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \sum_{i=1}^{card(B(\tau))} \int_{b^i} e^{\frac{||\boldsymbol{X}-\boldsymbol{Z}||^2}{2\sigma^2}} \boldsymbol{dZ}$$

$B(\tau)$ and $P_\Sigma(\tau)$ are computed thanks to a simple hierarchical algorithm based on the iterative increase of the partition depth (from $p_1 = 1$ to $p_p = p$). At each iteration, only the blocks for which the integral of the gaussian function is higher than τ are kept in a priority queue. Since $card(B(\tau))$ decreases with τ, finding B_α^{min} is equivalent to finding τ_{min} verifying:

$$\begin{array}{c} P_\Sigma(\tau_{min}) \geq \alpha \\ \forall \tau > \tau_{min}, P_\Sigma(\tau_{min}) < \alpha \end{array} \quad (6)$$

As $P_\Sigma(\tau)$ also decreases with τ, τ_{min} can be easily approximated by a method inspired by Newton-Raphson technique (the hierarchical algorithm is applied several times). The partition depth p is of major importance since it directly influences the total estimation time t_s:

$$t_s(p) = t_f(p) + t_r(p)$$

The time of the filtering step $t_f(p)$ is strictly increasing with p because the number of p-blocks in B_α^{min} and thus the computation time increase with p. The refinement time $t_r(p)$ is decreasing because the *selectivity* of the filtering step increases, i.e the number of features belonging to the selected blocks decreases with p. The search time $t_s(p)$ has generally only one minimum at p_{min} which can be set at the start of the system in order to obtain the best average response time. In practice, p_{min} depends particularly on the database size and varies from $p = 9$ to $p = 23$ when the database size grows from $5,000$ features to 1×10^8 features.

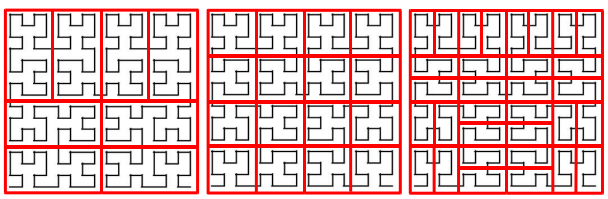

Figure 2: Space partition induced by the Hilbert space-filling curve at different depths $p = 3, 4, 5$

Once B_α^{min} has been determined, the physical address of the features belonging to each block is read in an index table and the refinement step is processed according to Equ. 4, N_α being the number of features belonging to B_α^{min}.

3. DENSITY-BASED SELECTION OF INTEREST POINTS

3.1 Local features and content-based image retrieval

The use of local features for content-based image retrieval (CBIR) was originally suggested by Schmid and al. [15, 12] and more recently applied to image copy detection by Berrani and al. [1] and to video copy detection in our previous work [7]. The extraction of local features consists of two steps: an interest point detection [11, 12, 6], and a local descriptor computation [13]. CBIR using local features argues that, instead of using a single feature vector to describe an entire image, one should identify and independently index a large number of local features. Instead of submitting a single query to retrieve similar images, multiple queries should be submitted and their partial results should be post-processed before delivering the answer.

Local features are well-suited to CBIR for two main reasons. First, they are ideal to deal with cropping, shifting and compositing because a part of them always remains unchanged, whereas a single global feature would need complex metrics to be robust to such transformations. Second, their local uniqueness and their high information content [16, 18] make them highly distinctive and robust to typical image transforms.

However, it is important to note that usual local features are selected only according to the local information content in the image. Thus, there is no guaranty that they will be distinctive in a large set of local features. A local feature corresponding to a high saliency in the image could be highly redundant in some specific databases.

To overcome this issue, we propose to select relevant local features directly according to their discrimination power in a specific set of images. By computing the density of the local features in a source database (with our proposed method), it is indeed possible to select quickly the most *rare* local features. Such a rarity criterion was already proposed in the literature to select salient points [19, 5] but the source features to estimate the density were only the local features of the current image or a small class of similar images [19]. In our method the rarity of a local feature is related to a large set of various images and thus more specific to CBIR issues.

3.2 Density-based selection of discriminant Harris interest points

Although the density-based selection of interest points could be processed on all the pixels of an image, we restricted it to the post-selection of Harris interest points [6]. The Harris detector works as a first filtering step, and the density is computed only for the local features extracted around the detected points. We finally keep only the K_r local features with the lowest density. The source database used to estimate the density is a large set of local features extracted around all Harris interest points of a represen-

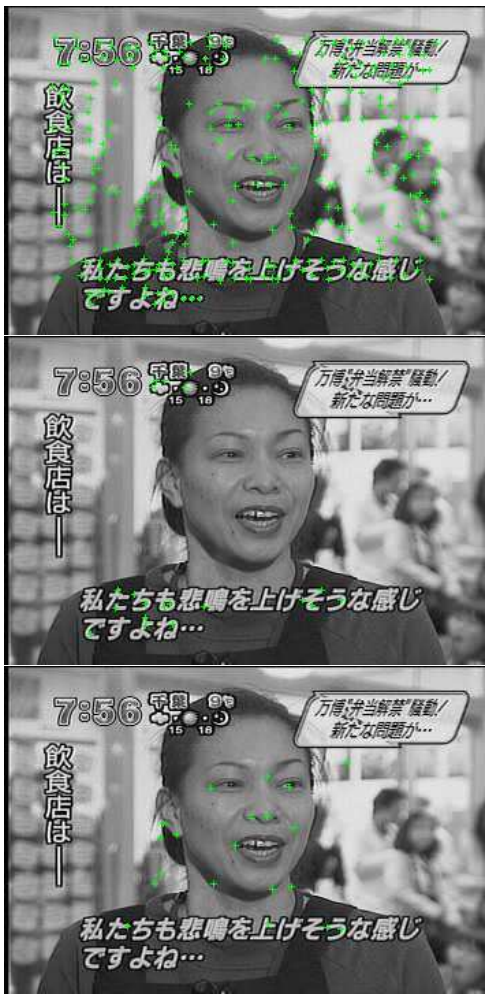

Figure 3: top: all harris points - center: 20 points with the highest harris response - bottom: 20 points with the lowest density

tative collection of object (depending on the application). Fig. 3 illustrates the advantage of our density-based selection criterion compared to a *cornerness* selection criterion. The top image represents all the detected Harris points, the second image represents the $K_h = 20$ points with the highest Harris response (i.e the highest cornerness) and the bottom image represents the $K_r = 20$ points with the lowest density in a source database containing the same kind of images (Japanese TV programs). We see that the points with the highest cornerness correspond only to textual characters which are very frequent in the database and therefore not discriminant at all. On the other side, the points with the lowest density focus more on the main information of the scene.

4. EXPERIMENTS

4.1 Experimental setup

All the features used in the following experiments are local descriptors computed around Harris interest points [6, 7] in video clips. The video materials come from Japanese television channels that are saved in MPEG1 format for 5 years. They include TV shows, news, movies, sports, etc. The feature extraction method was used in [7] for content-based video copy detection. It includes a key-image detection (corresponding to extrema of the global intensity of motion [2]), the detection of Harris interest points [6] in these key-images and the computation of 20-dimensional local features defined as:

$$\boldsymbol{S} = \frac{\boldsymbol{s^1}}{\|\boldsymbol{s^1}\|}, \frac{\boldsymbol{s^2}}{\|\boldsymbol{s^2}\|}, \frac{\boldsymbol{s^3}}{\|\boldsymbol{s^3}\|}, \frac{\boldsymbol{s^4}}{\|\boldsymbol{s^4}\|}$$

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

$$\boldsymbol{s^i} = \frac{\partial \boldsymbol{I}}{\partial x}, \frac{\partial \boldsymbol{I}}{\partial y}, \frac{\partial^2 \boldsymbol{I}}{\partial x \partial y}, \frac{\partial^2 \boldsymbol{I}}{\partial x^2}, \frac{\partial^2 \boldsymbol{I}}{\partial y^2}$$

All measurements and parameters refer to the normalized feature space $[0, 1]^{20}$. The average key-image rate is about 0.95 key-image per second of video and the average number of Harris interest points per key image is about 170.
Experiments were carried out on a Pentium M (CPU 1.86 GHz, cache size 2048Kb, RAM 1.5 Gb) and the response times were obtained with unix `getrusage()` command.

4.2 Fast Density Estimation Technique evaluation

We aim at evaluating our density estimation technique in terms of quality and speed. In this experiment the feature databases were constructed by indexing the differential descriptors of all the harris interest points detected in the key-images. The default database DB_{10}, used when the size of the source dataset is not a studied parameter, contains $5,814,585$ feature vectors corresponding to 10 hours of randomly selected video materials. Two error metrics are used to assess the quality of the density estimation:

- the relative error:

$$\eta = \frac{\widehat{f}(\boldsymbol{X}) - \widehat{f}_\alpha(\boldsymbol{X})}{\widehat{f}(\boldsymbol{X})}$$

- the relative logarithmic error:

$$\eta_{log} = \frac{log_{10}(\widehat{f}(\boldsymbol{X})) - log_{10}(\widehat{f}_\alpha(\boldsymbol{X}))}{log_{10}(\widehat{f}(\boldsymbol{X}))}$$

The second one is more adapted to the dynamic of probability density functions in high dimensional spaces (see Fig. 4 for example).
The default bandwidth is set to $\sigma = 0.117$.
Note that the target features for which we estimate the density do not belong to the source databases.

4.2.1 *Influence of the precision parameter* α

The estimated density of 30 local features extracted in a randomly selected key-image is represented in Fig. 4 for several values of α. The reference value estimated by a sequential scan of the source database is also plotted on each graphic. This qualitative analysis show that the quality of the estimation remains almost unchanged for high precision values ($\alpha \geq 99\%$) and that it begins to seriously degrade only for precision values lower than $\alpha = 80\%$.

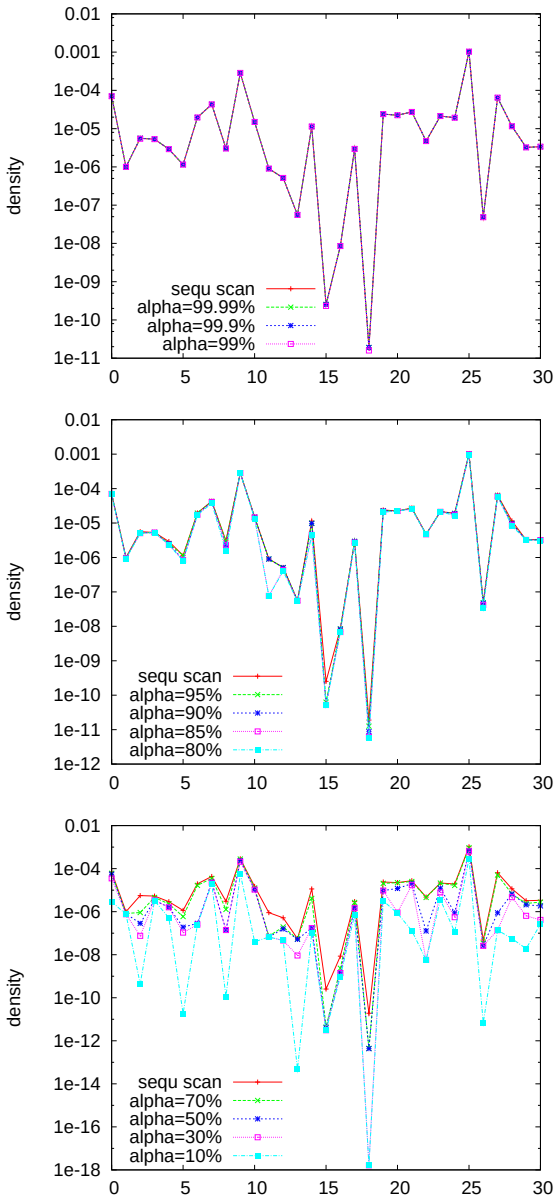

Figure 4: Estimated density of 30 local features, for several values of the precision parameter α

Fig. 5 shows the average estimation time of one single target feature faced with the value of α. It shows that the approximate estimation paradigm allows very high speed-up with only small losses in quality thanks to the strong decrease of the estimation time when the precision is slowly decreasing (the estimation time is divided by 21 when α varies from 99.99% to 95%).
Speed measurements and quantitative quality measurements are summarized in Table 1. For most computer vision application, using $\alpha = 90\%$ should provide a widely acceptable precision ($\eta_{log} < 1\%$) whereas the estimation is about 100 times faster than using a sequential scan of the source database.

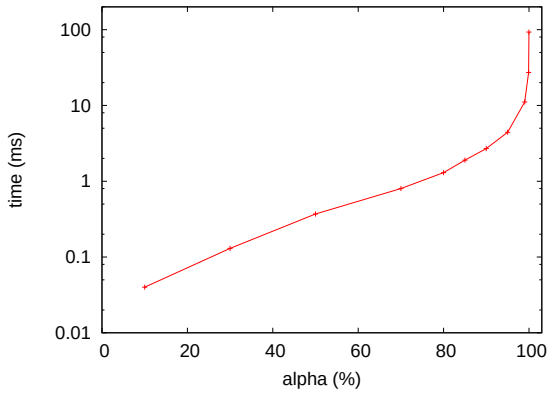

Figure 5: Estimation time versus α (log scale)

Table 1: Influence of the precision α

	time (ms)		
sequential scan	266.27		
α (%)	time (ms)	η (%)	η_{log} (%)
99.99	93.31	0.0005	0.0001
99.90	27.33	0.0083	0.0007
99.00	11.17	0.5099	0.0331
95.00	4.40	4.1921	0.3541
90.00	2.71	9.8760	0.9458
85.00	1.91	15.0217	1.7295
80.00	1.29	19.7469	2.3080
70.00	0.84	30.8985	4.5035
50.00	0.37	51.5602	11.2620
30.00	0.13	70.9108	23.2837
10.00	0.04	85.1596	39.6444

4.2.2 *Influence of the bandwidth σ*

Table 2: Influence of the bandwidth σ

	time (ms)		
sequential scan	266.27		
σ	time (ms)	η (%)	η_{log} (%)
0.05	0.61	34.42	5.76
0.10	2.14	12.31	1.30
0.15	4.07	6.65	0.93
0.20	7.83	5.72	0.91

Table 2 summarizes speed and quality measurements for several values of σ ($\alpha = 90\%$). It shows that the bandwidth has an important impact on the evaluation time. We can also remark that the quality degrades for very low values ($\sigma = 0.05$). This is highly related to the floating precision of the machine that plays a more important role when the values of the density are very low.

4.2.3 *Influence of the source database size*

Fig. 6 represents the average estimation time of one single target feature for several sizes of the source database. The main advantage of our technique is that it is sublinear in database size when it remains of the same order of magnitude as the memory size of the machine. In this experimental context, the size of an indexed database containing

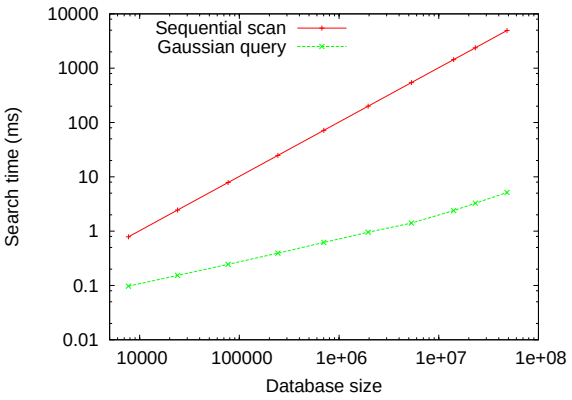

Figure 6: Estimation time versus database size (log scale)

$50,000,000$ features is about $1.5Gb$. For such a database, our technique is about 1000 times faster than a sequential scan.

4.3 Discriminant interest points selection

We aim at evaluating the density estimation of local features as a criterion to select relevant interest points. The principle of the selection is to extract the local features around each Harris interest points, to estimate their density in a source database using our fast density estimation technique and to keep only those having the lowest density (most *rare* points). The source database we use is the default database DB_{10}, previously described in section 4.2, and containing the local features of all the Harris points detected in 10 hours of randomly selected Japanese television programs. Note that the video materials in which we perform the interest points selection do not belong to the 10 hours of video used to construct DB_{10} although they are of same type (Japanese television programs).

4.3.1 Qualitative analysis

Fig. 7 and 8 represent the $K_h = 20$ Harris interest points with the highest cornerness (i.e the highest Harris response) and the $K_r = 20$ Harris interest points with the lowest density in several key-images. These examples show that the density criterion is most relevant than the cornerness criterion to select discriminant points. The Harris points with the highest cornerness are often detected in the background of the video or in inserted patterns (text, logos, frames,...). Thus, they are not discriminant to distinguish different scenes of the same video clip or different programs with the same background or patterns. The rare local features are better distributed in the image and capture a more representative information of the scene.

Fig. 8 shows how the points with the lowest densities are well suited to capture motion information. The local features describing static local regions are indeed very frequent in the source database (background, static scenes, etc.) and the most rare local features often correspond to *space-time* interest points [10].

4.3.2 Content-based copy detection

To perform a quantitative analysis of the discrimination power of low density local features, we use them in the

Figure 7: top: 20 points with the highest harris response - bottom: 20 points with the lowest density

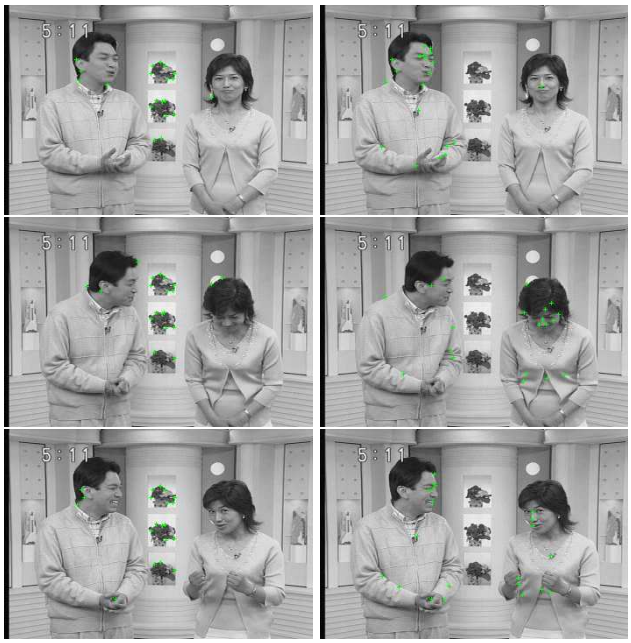

Figure 8: left: 20 points with the highest harris response - right: 20 points with the lowest density

content-based copy detection framework (CBCD) described in [7]. In this previous work the detection was based on the local features extracted around the $K_h = 20$ Harris points with the highest cornerness in each key-image. We propose here to replace these features by the $K_r = 20$ Harris points with the lowest density in a source database and to compare the results. The principle of our experiment is the following: the local features (either the most rare or with the most cor-

nerness) are first computed from the video clips belonging to the reference catalog and they are inserted in an indexing structure (similar to the one we use to perform fast density estimation). The local features are then computed in each candidate video clip and the similar local features are searched in the indexed database. Finally, a post-processing vote is applied to the partial results to decide which reference video clips are copies of the candidate video clip.
The reference catalog is composed of about 150 hours of Japanese TV programs represented by $10,254,372$ local features in the indexed database. The source database used to determine the density of the local features is the previously described database DB_{10} (containing the local features of all the Harris points in 10 hours of randomly selected Japanese television programs, see section 4.1 and 4.2). Note that both candidate and reference video clips do not belong to the video set used to construct the source database DB_{10}.

1. **Speed performances:**
   Table 3 summarizes the speed measurements for both methods ($HARRIS_{max}$ refers to the old method that keeps only the 20 interest points with the highest harris response and $HARRIS_{rare}$ to the new method that keeps only the 20 interest points with the lowest density). T_{index} is the total time to index the 150 hours of video (including local features extraction and insertion in the indexing structure). T_{detect} is the total average time to process 1 hour of candidate video clips (including local features extraction time $t_{extract}$, density estimation time $t_{density}$, search time t_{search} and post-processing time t_{vote}).

Table 3: Speed measurements

Method	$HARRIS_{max}$	$HARRIS_{rare}$
T_{index} (s)	68,737	144,003
T_{detect} (s)	658	983
$t_{extract}$ (s)	447	449
$t_{density}$ (s)		453
t_{search} (s)	33	13
t_{vote} (s)	178	68

   We see that the total time to extract the 20 most rare features ($t_{extract}+t_{density}$) is only about 2 times slower than the time to extract the points with the highest Harris response. Thanks to our fast density density estimation technique the density estimation step is about 8 times faster than real time ($453s$ to process $3600s$). The indexing time which mainly depends on the feature extraction time is also about 2 times slower.
   On the other hand, the total search process ($t_{search} + t_{vote}$) is 2.60 times faster when using the most rare points. This can be explained by the higher discrimination of these points that reduces the number of comparisons during the search and the number of neighbors in the partial results that need to be post-processed. For example, the average number of neighbors retrieved in a range query of radius $r = 0.35$ is equal to 2.57 for the most rare points and to 6.05 for the points with the highest Harris response.

2. **Duplicates detection:** In this experiment, we simply searched 12 hours of TV programs belonging to the reference catalog in order to detect duplicates (for instance video clips used in different news, replicated broadcasts, etc.). The results were manually controlled for both methods and we count the number of good detections and false alarms. The results are summarized in table 4 and clearly show the improvement of using the most rare local features since the number of good detections is almost the same (and even better for the new method) whereas the precision increases from 76.34% to 95.53%.

Table 4: Duplicates detection results

Method	$HARRIS_{max}$	$HARRIS_{rare}$
number of detections	334	269
number of true detections	255	257
number of false detections	79	12

Figure 9: Examples of false positives rejected by the new technique

Figure 10: Examples of a true positive detected only by the new technique

   Two false positive matches rejected by the new method are displayed on Fig. 9 and one good match detected only by the new method is displayed on Fig. 10.

3. **Recall/Precision curves:** We built a synthetic ground truth in order to construct Recall/Precision curves for both methods. The true probes were obtained by processing several video clips randomly selected in the reference catalog with several transformations or combinations of transformation (including gamma and contrast modification, encoding/decoding, small resizing, shifting, gaussian noise addition and texts insertion). The false probes were obtained by selecting programs that are not included in the reference catalog and by controlling manually that the most confidence matches were false alarms. Note that the total length of the

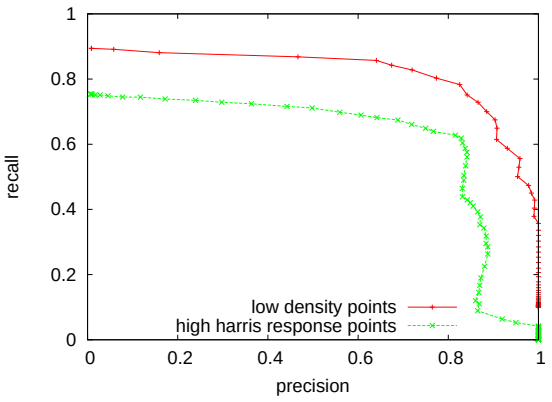

Figure 11: ROC curves of the copy detection

false probes is 10 times longer than the length of the true probes in order to be more realistic. The resulting curves are displayed on Fig. 11 and show clearly that the use of the most rare local features improve significantly the retrieval performances of the system. At constant precision, for instance, the recall improvement raises from 15% to 50% for high precision values.

5. CONCLUSIONS

In this paper, we propose a fast approximate density estimation technique based on the energy of the gaussian function in a simple grid index structure. This method enables the density estimation of more than 200 target features per second with a very large source features dataset (until 100 millions source features). This makes possible the real time density estimation of local features extracted in a video stream and we have shown how it can be used to enhance significantly the performances of a content-based copy detection framework by selecting more discriminant interest points. We think that this kind of selection criterion, computed according to the distribution in the database and not only to the information contained in the image itself is a promising direction to select features relevant for indexing purposes. In future work, we will attempt to select directly interest points by their density without using the previous Harris detection step. The evolution of the density when varying the bandwidth of the estimation is also an interesting investigation track to select relevant features.

6. ACKNOWLEDGMENTS

The authors would like to thank the Japanese National Institute of Informatics who generously provided the video clips used in this paper and especially Pr. Satoh and F. Yamagishi who helped us to perform the experiments.

7. 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*, pages 70–77, 2003.

[2] 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.

[3] A. G. Gray and A. W. Moore. Nonparametric density estimation: Toward computational tractability. In *Proc. of Int. Conf. on Data Mining*, 2003.

[4] L. Greengard and J. Strain. The fast gauss transform. *SIAM J. Sci. Stat. Comput.*, 12(1):79–94, 1991.

[5] D. Hall, B. Leibe, and B. Schiele. Saliency of interest points under scale changes. In *roc. of the British Machine Vision Conference*, 2002.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of Alvey Vision Conf.*, pages 147–151, 1988.

[7] 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*, pages 414–424, 2003.

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

[9] A. Joly, C. Frélicot, and O. Buisson. Statistical similarity search applied to content-based video copy detection. In *IEEE Int. Workshop on Managing Data for Emerging Multimedia Applications*, 2005.

[10] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. of Int. Conf. on Computer Vision*, pages 432–439, 2003.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of Int. Conf. on Computer Vision*, pages 1150–1157, 1999.

[12] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. of Int. Conf. on Computer Vision*, pages 525–531, 2001.

[13] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Submitted to Trans. on Patter Ananlysis and Machine Intelligence*, 2004.

[14] A. W. Moore. The anchors hierarchy: Using the triangle inequality to survive high dimensional data. In *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 397–405, 2000.

[15] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[16] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Proc. of Int. Conf. on Computer Vision*, pages 230–235, 1998.

[17] D. W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1992.

[18] N. Sebe and M. S. Lew. Comparing salient point detectors. *Pattern Recognition Letters*, 24(1):89–96, 2003.

[19] K. N. Walker, T. F. Cootes, and C. J. Taylor. Locating salient object features. In *Proc. of the British Machine Vision Conference*, 1998.