A general principled method for image similarity validation

Frédéric Cao and Patrick Bouthemy

IRISA / INRIA
Campus universitaire de Beaulieu
35042 Rennes Cedex, France
{fcao,bouthemy}@irisa.fr

Abstract. A novel and general criterion for image similarity validation is introduced using the so-called *a contrario* decision framework. It is mathematically proved that it is possible to compute a fully automatic detection criterion to decide that two images have a common cause, which can be taken as a definition of similarity. Analytical estimates of the necessary and sufficient number of sample points are also given. An implementation of this criterion is designed exploiting the comparison of grey level gradient direction at randomly sampled points. Similar images are detected *a contrario*, by rejecting an hypothesis that resemblance is due to randomness, which is far more easy to model than a realistic degradation process. The method proves very robust to noise, transparency and partial occlusion. It is also invariant to contrast change and can accomodate global geometric transformations. It does not require any feature matching step. It can be global or local, only the global version is investigated in this paper.

1 Introduction

Establishing that two images, or parts of images, are similar is a general concern in image analysis and computer vision. It is involved in a number of problems or applications, and more specifically in image or video retrieval [1, 16]. In this paper, we answer the following question: can we automatically assess that two images are similar and with which degree of confidence? A second question is: can we compute "universal" thresholds to decide that two images are similar? This problem is very difficult in full generality since image similarity should be defined up to a large group of invariance, which may depend on the application: contrast change, occlusion, transparency, noise, translation, scaling, geometric deformation, etc.
In this paper, we investigate the global case. Even on complete images, this is a central issue for image retrieval: checking whether or not an image is present in a database or in a video stream. The designed solution is based on statistical arguments. It requires very simple information computed on the image intensities. It is extremely stable with respect to noise (it still works with an additive Gaussian noise with standard deviation 30 or a 50% impulse noise). The search is

totally processed online and is very efficient (10 frames/s on a 2.4GHz PC, with no optimization). The implemented version only relies on the direction of the image gradient, and is therefore contrast invariant. We have demonstrated that it is robust to occlusion and transparency. Finally, we will mention how global geometric transformations can be handled. Let us point out that the similarity measure does not require any feature matching step.

The paper is organized as follows. A brief review of related work is made in Section 2. In Section 3, the *a contrario* decision framework is introduced and used to define an automatic criterion for the similarity between two images. The method will be introduced in parallel of a more usual hypothesis testing framework, but we emphasize that decision only relies on the likelihood of one hypothesis (which is that the two compared images are not the same). The implemented test compares the image gradient direction at some random points. Similarity is detected *a contrario*, by rejecting an hypothesis that resemblance is due to randomness. In Section 4, we show that this number of sample points can be chosen to maintain a probability of detection very close to 1, when we assume white Gaussian noise. However, we insist that detection does not rely on such a Gaussian noise assumption. It will be observed that, in practice, the required number of samples is seldom above a few hundreds, even for quite important noise. Section 5 contains experimental results of image comparison and retrieval in databases of typically $10,000$ images. We cope with several kinds of image perturbations as strong Gaussian and impulse noise, JPEG compression, transparency, occlusion. We also handle a prior registration before detecting similar pairs. Summary and concluding remarks are given in Section 6.

2 Related work

The statistical arguments we introduce can be related to the work of Lisani and Morel [8]. Their approach uses the direction of the gradient of a grey level image, and they detect local changes in registered stereo pairs of satellite images. Our method is dual since, on the contrary, we use the gradient direction in both images to decide that they have much spatial information in common. Detection thresholds are computed by using an *a contrario framework*, as introduced by Desolneux, Moisan and Morel [2], and extended for spatio-temporal problems in [18]. More ancient work [17] used the same kind of ideas but detection thresholds were not computed. Other image features widely used are SIFT descriptors [9, 10] which are basically local direction distributions. Nevertheless, the indexing and comparison of descriptors is achieved by a nearest-neighbor procedure. Hence, there is no decision involving an automatic threshold setting, which is precisely our main concern. On the other hand, we think that our methodology can be adapted to the comparison of SIFT features as well, instead of using the direction of the spatial intensity gradient.

Basically, our method consists in sampling random points in two images and counting the number of points such that the difference of the spatial intensity gradient direction is small enough. Using the gradient direction as image feature

for image similarity detection was already proven useful (e.g., [13]). This step is embedded in a probabilistic framework which will be subsequently discussed. Let us point out that contrarily to methods as RANSAC [4], the estimation of the registration parameters is completely separated from the similarity decision step, which makes the proposed method more general. In particular, our method can consider different types of image features, independently of the image information used to perform the registration. Furthermore, it can be used to validate the performance of the registration methods themselves.

Probabilities will be computed in a model representing the absence of similarity (so-called *background model*, in the statistical meaning). Some similar idea can be found in [5] where the authors study the influence of "conspiracy of random".

3 A contrast invariant image comparison method

In what follows, we always assume that images are grey-level valued with size $N \times N$. Let u and v be two images. To facilitate understanding, the development below is instantiated for the case where image gradient direction is the considered image feature. However, let us stress that this framework is general and other kinds of image features could be utilized as well.

For any point x, let us denote by $\theta_u(x)$ and $\theta_v(x)$ the directions of the image gradient of u and v at point x. Let us denote by $D_{u,v}(x)$ the angle difference between $\theta_u(x)$ and $\theta_v(x)$ on the unit circle $\mathbb{S}^1$. When there is no risk of ambiguity, we elude the subscript and write $D(x)$ instead. It is a real value in $[0, \pi]$. Since we want this measure to be accurate, we only consider points where both image gradients are large enough (larger than 5 in practice). Now, two images differing from a contrast change have the same gradient direction everywhere, which ensures that the method is contrast invariant.

Even though the proposed method is not a classical hypothesis testing, let us formulate it this way, to explain its principle. From the observations of the values of $D(x)$, let us consider that we aim at selecting one of the two following hypotheses: $\mathcal{H}_0$: u and v are unrelated images. $\mathcal{H}_1$: u and v have similar content. Modeling Hypothesis $\mathcal{H}_1$ is equivalent to model the type of degradation that can lead from u to v, and only very simplistic models are usually at hand. In an image retrieval application, v can belong to a database of typically 10^6 images (10 hours of video). Hence, false alarms (that is, accept $\mathcal{H}_1$ while $\mathcal{H}_0$ actually holds) have to be controlled, else the system will become impractical. Because of the large size of the database, this implies that it is necessary to ensure very small probabilities of false alarms. The proposed method is to base the decision only on $\mathcal{H}_0$, which is far more easy to model. It allows us to attain very small probabilities of false alarm. Moreover, there is no need to compare the likelihood of the two hypotheses, since we can derive automatic thresholds on the likelihood of $\mathcal{H}_0$, which allows us to reject it very surely.

Hypothesis $\mathcal{H}_0$ models the absence of similarity. Thus, the following assumption is made: for some set of points x_1, ..., x_M, the values $D(x_i)_{i \in \{1,\dots,M\}}$ are

independent, identically distributed in $[0, \pi]$. This probabilistic model will be called the *a contrario* model (or background model). The principle of the detection is to compute the probability that the real observation has been generated by the *a contrario* model. When this probability is too small, the independence assumption of the two images is rejected and similarity is detected (validated).

Let $\alpha \in (0, \pi)$, and $q_\alpha = \frac{\alpha}{\pi}$ be the probability that the considered angle is less than or equal to α. For any set of distinct points $\{x_1, \ldots, x_M\}$, the probability, under $\mathcal{H}_0$, that at least k among the M values $\{D(x_1), \ldots D(x_M)\}$ are less than α is given by the tail of the binomial law

$$B(M, k, q_\alpha) = \sum_{j=k}^{M} \binom{M}{j} q_\alpha^j (1 - q_\alpha)^{M-j}.$$

Definition 1. *Let $0 \leqslant \alpha_1 \leqslant \ldots \leqslant \alpha_L \leqslant \pi$ be L values in $[0, \pi]$. Let u a real valued image, and $x_1, \ldots x_M$, M distinct points. Let us also consider a database $\mathcal{B}$ of $N_\mathcal{B}$ images. For any $v \in \mathcal{B}$, we call number of false alarms of (u, v) the quantity*

$$NFA(u, v) = N_\mathcal{B} \cdot L \cdot \min_{1 \leqslant i \leqslant L} B(M, k_i, q_{\alpha_i}), \tag{1}$$

where k_i is the cardinality of

$$\{j,\, 1 \leqslant j \leqslant M, D_{u,v}(x_j) \leqslant \alpha_i\}.$$

We say that the pair (u, v) is meaningful (more specifically, ε-meaningful), or that u and v are similar (more specifically, ε-similar) if $NFA(u, v) \leqslant \varepsilon$.

The interpretation of this definition will be made clear after stating the following proposition. Let us just mention now that the probability given by the tail of the binomial law has to be multiplied by the number of tests done, i.e., the considered number (L) of quantized values of the gradient direction and the overall number ($N_\mathcal{B}$) of tested images, to evaluate the NFA.

Proposition 1. *For a database of $N_\mathcal{B}$ images such that the gradient direction difference with a query u has been generated by the background model, the expected number of v such that (u, v) is ε-meaningful is less or equal than ε.*

Proof. For all i, let us denote by K_i the random number of points among the x_j such that $D(x_j)$ is less than α_i. For any v, (u, v) is ε-meaningful, if there is at least $1 \leqslant i \leqslant L$ such that $N_\mathcal{B} \cdot L \cdot B(M, K_i, q_{\alpha_i}) < \varepsilon$. Let us denote by $E(v, i)$ this event. Its probability $P_{\mathcal{H}_0}(E(v, i))$ satisfies

$$P_{\mathcal{H}_0}(E(v, i)) \leqslant \frac{\varepsilon}{L \cdot N_\mathcal{B}}.$$

Indeed, for any real random variable X with survival function $H(x) = P(X > x)$, it is a classical fact that $P(H(X) < x) \leqslant x$. By applying this result to K_i, we

get the upper bound on $P(E(v,i))$. The event $E(v)$ defined by "(u,v) is ε-meaningful" is $E(v) = \cup_{1 \leqslant i \leqslant L} E(v,i)$. Let us denote by $\mathbb{E}_{\mathcal{H}_0}$ the mathematical expectation under the *a contrario* assumption. Then

$$\begin{aligned}\mathbb{E}_{\mathcal{H}_0}\left(\sum_{v\in\mathcal{B}} \mathbf{1}_{E(v)}\right) &= \sum_{v\in\mathcal{B}} \mathbb{E}_{\mathcal{H}_0}(\mathbf{1}_{E(v)}) \\ &\leqslant \sum_{\substack{v\in\mathcal{B}\\ 1\leqslant i\leqslant L}} P_{\mathcal{H}_0}(E(v,i)) \\ &\leqslant \sum_{\substack{v\in\mathcal{B}\\ 1\leqslant i\leqslant L}} \frac{\varepsilon}{L\cdot N_{\mathcal{B}}} = \varepsilon. \qquad \square\end{aligned}$$

Definition 1 together with Proposition 1 mean that there is in average less than ε images v in the database $\mathcal{B}$ that could match with u by chance, that is to say, when $\mathcal{H}_0$ holds. As a matter of fact, any detection must be considered as a false alarm under hypothesis $\mathcal{H}_0$ (hence the denomination of NFA - number of false alarms -, which might be at first misleading for the reader since the NFA value is used to detect the really similar image pairs, as specified in the Algorithm summary given next page).

Thus, it is chosen to eliminate any observation (i.e., any image v, given image u) having a frequency of the order of ε (or more) in the *a contrario model*. In Section 5.1, it will be checked that Hypothesis $\mathcal{H}_0$ is sound for two unrelated images.

Even though this is theoretically simple, it may be difficult to numerically evaluate the tail of the binomial law. A sufficient and more tractable condition of meaningfulness is given by the following classical result, first proved by Hoeffding [6].

Proposition 2. *Let $H(r,p) = r\ln\frac{r}{p} + (1-r)\ln\frac{1-r}{1-p}$, be the relative entropy of two Bernoulli laws with parameters r and p. Then, for $k \geqslant Mp$,*

$$B(M,k,p) \leqslant \exp\left(-M\cdot H\left(\frac{k}{M},p\right)\right). \tag{2}$$

This inequality leads to the following sufficient condition of meaningfulness.

Corollary 1. *If*

$$\max_{\substack{1\leqslant i\leqslant L\\ k_i\geqslant Mq_{\alpha_i}}} H\left(\frac{k_i}{M},q_{\alpha_i}\right) > \frac{1}{M}\ln\frac{LN_{\mathcal{B}}}{\varepsilon}, \tag{3}$$

the pair (u,v) is ε-meaningful.

In this corollary, it appears clearly that the values of k such that (u,v) is ε-meaningful only depends on the logarithm of L, $N_{\mathcal{B}}$ and ε. In practice, we choose L about 32 which is compatible with our perceptual accuracy of directions. In

other terms, the α_i must be understood as quantization steps of $(0, \pi)$. We also take $\varepsilon = 1$ since it means that we may have in average less than 1-false detection. However, as we shall see, really similar images have much smaller NFA and the choice of ε is not really important. Thus, in all experiments, we always set $\varepsilon = 1$, and we can therefore claim that the decision threshold is automatically derived.

The algorithm to be implemented is actually simple and of very low computational complexity. Indeed, it involves only a few computations as indicated below.

Algorithm

Let us fix $M > 1$, and L quantized values $(\alpha_i)_{1 \leqslant i \leqslant L}$.
For a pair of image u, v:

1. Draw M random points x_1, ..., x_M.
2. Compute the difference of the gradient direction $D(x_j)$.
3. For each i
   (a) Count the number of x_j such that $D(x_j) \leqslant \alpha_i$, denoted by k_i.
   (b) Compute $N_{\mathcal{B}} \sum_{n=k_i}^{M} \binom{M}{n} q_{\alpha_i}^n (1 - q_{\alpha_i})^{M-n}$ (with $q_{\alpha_i} = \frac{\alpha_i}{\pi}$).
4. $NFA(u, v)$ is the minimum of these values.
5. Test if $NFA(u, v) \leqslant \varepsilon$.

In practice, we take M varies between 200 and 500 (this is discussed below), $\alpha = 32$ (this hardly has any incidence). Let us point out that the quantity $-\log_{10} NFA$ can be considered as a confidence level, while being a more tractable number.

4 Random sampling

4.1 Problem statement

The *a contrario* model assumes that the values $D(x_j)$ are i.i.d. in $(0, \pi)$. This implicitly means that it is assumed that the direction $\theta_u(x_j)$ and $\theta_v(x_j)$ are independent for a given x_j, and that all the directions $\theta_u(x_j)$ are also mutually independent. (The same holds for v.) The NFA is nothing but a measure of the deviation to this hypothesis. If a few points are randomly drawn in the image, this assumption is clearly reasonable. However, since natural images contain alignments the second assumption becomes clearly false if we sample too many points. Moreover, if the two images have a casual alignment in common, this segment will induce a very strong deviation from the independence assumption, and the images could be wrongly considered as similar. We then face the following dilemma for choosing the number of samples M:

- it must be large enough to allow us to contradict the independence hypothesis and to obtain small values of the number of false alarms for two similar images.

– it must be small enough to avoid the "common alignment problem". If we draw a few hundreds points uniformly in the images, then they are aligned very unlikely.

In order to evaluate the typical magnitude of the number of sample points, let us assume that v differs from u by an additive Gaussian noise $\mathcal{N}(0, \sigma^2)$, which will be our hypothesis $\mathcal{H}_1$. We insist that we use this $\mathcal{H}_1$ to only determine the magnitude of the sufficient number of sample points, but since we cannot assert that this model is realistic, the detection eventually relies only upon the background model $\mathcal{H}_0$. By computing the gradient by a finite difference scheme, it is possible to assume that the gradient coordinates of v are also corrupted by a white Gaussian noise (with a variance depending on the numerical scheme). If the law of the gradient norm is empirically estimated, it becomes possible to compute the law of the direction variation D, $P_{\mathcal{H}_1}(D < \alpha)$.

4.2 Bounds on the number of sample points

By definition, we detect the pair (u, v) as ε-meaningful, if $NFA(u, v) < \varepsilon$. If $\mathcal{H}_1$ holds, we would like to detected meaningful pairs with a high probability. Hence, we would like the value $P(NFA(u, v) < \varepsilon | \mathcal{H}_1)$ to be large whenever v is a (noisy) version of u. Let us also assume that u is an image of a query base $\mathcal{Q}$ containing $N_{\mathcal{Q}}$ images (and v is still in the database $\mathcal{B}$). If we want less than ε detection in the *a contrario* model by comparing all the pairs in $\mathcal{Q} \times \mathcal{B}$, we have to multiply the NFA definition (1) by $N_{\mathcal{Q}}$. Let

$$k_\alpha = \inf\{k, \text{ s.t. } N_{\mathcal{Q}} \cdot N_{\mathcal{B}} \cdot L \cdot B(M, k, q_\alpha) < \varepsilon\}.$$

To make things simpler, assume that we compute the NFA with only one value of angle α (so that $L = 1$). Since there is no ambiguity, we drop the subscript α. If K is the random number of points such that $D < \alpha$, the pair (u, v) is detected if and only if $K \geqslant k$. The probability of detection under $\mathcal{H}_1$ is therefore

$$P_D \equiv P(K \geqslant k | \mathcal{H}_1) = B(M, k, p). \tag{4}$$

where

$$p = P_{\mathcal{H}_1}(D < \alpha),$$

which is known, since we have here a model of noise.

Definition 2. *We call number of misses*

$$\mathcal{M}(M, k) = N_{\mathcal{Q}} N_{\mathcal{B}} (1 - B(M, k, p)). \tag{5}$$

As for the number of false alarms, if $\mathcal{M}(M, k) < \varepsilon$, it is clear that the expected number of misdetections under hypothesis $\mathcal{H}_1$ is less than ε.

The noise model clearly implies that p (the probability that gradient directions are alike when both images are the same) is larger than q (probability that the directions are alike for unrelated images, i.e. the *a contrario* model) unless

the images are constant of $\sigma = +\infty$, which is of little interest, and $p \to q$ when $\sigma \to +\infty$ (up to a normalization of grey level, the image tends to a white noise).

From estimates on the tail of the binomial law, we obtain the following necessary conditions on the number of samples M.

Proposition 3. *Assume that $\mathcal{M}(M,k) < \varepsilon$. Then, for some positive constant $C \simeq 0.39246$,*

$$M(p-q)^2 \geqslant \min(p(1-p), q(1-q)) \left(C + \ln \frac{N_{\mathcal{Q}} N_{\mathcal{B}}}{\varepsilon \sqrt{M}} \right). \tag{6}$$

The proof is given in appendix.

The estimate above tells that, when the noise amount σ becomes large, M grows like $\frac{1}{(p-q)^2}$. This is not strictly exact because of the $\ln M$ term on the right side of (6). This term is unavoidable since it appears in any sharp lower bound of the tail of the binomial law. In the following Proposition 4, it will be proved that the order of magnitude $O((p-q)^{-2})$ is sufficient.

Proposition 4. *If*

$$M \geqslant \frac{2}{(p-q)^2} \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon}. \tag{7}$$

then $\mathcal{M}(M,k) < \varepsilon$.

In practice, we do not know neither that the two images are the same nor the amount of noise. However, the purpose of this result is to determine the order of magnitude of the sufficient number of sample points. Numerical evaluation shows that it is a few hundreds which is compatible with the size of usual images.

5 Numerical applications and experiments

5.1 Justification of the background model

The background model should be sound for two unrelated images. Let us make the following experiment. Let us compute the empirical distribution of the gradient direction on two images. Because of quantization and presence of strongly privileged directions, these two histograms are not uniform at all. Nevertheless, the distribution of the difference of the directions, taken at *two* random locations (that is, different points in the two images) is the circular convolution of these histograms. On many pairs of images, we indeed checked that the difference of the repartition function with a uniform distribution in $(-\pi, \pi)$ is everywhere less than 0.01.

5.2 Number of sample points under hypothesis $\mathcal{H}_1$

On Fig. 1, we discuss (see the caption) the relation between σ (the noise standard deviation), M (the number of sample points) and the detection rate as explained in subsection 4.2. By varying σ and M, we empirically retrieve the bound estimate of subsection 4.2.

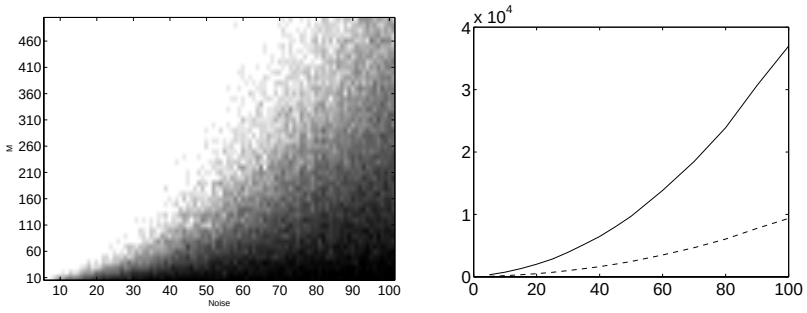

Fig. 1. We match an image with some of its corrupted versions by a white Gaussian noise, for σ varying between 5 and 100 (horizontal axis), and for a number of samples M between 10 and 500 (vertical axis). For each couple (σ, M), 50 trials are drawn, yielding $N_{\mathcal{B}} = 250000$. The grey level in the left plot is the number of similarity detections (white for 50 and black for 0). The curves on the right are the sufficient and necessary values of M for controlling the number of misses, given by (6) and (7) respectively. As expected, the empirical results on the left are between these curves and bounds are not sharp.

5.3 Experiments of image retrieval and image comparison

We have tested the robustness of the method for image retrieval in a video stream with respect to the following degradations: noise (impulse, Gaussian or JPEG compression), transparency, partial occlusion. The image comparison is directly applied with no preprocessing of any type. There are actually some applications to such a detection method: for instance, to segment television video stream one may look for particular jingles or some recurrent images. Current methods work by computing local features and matching them. It thus requires to pre-compute those features, organize and store them in feature databases. The proposed method only needs the spatial image gradient on a few hundred points.

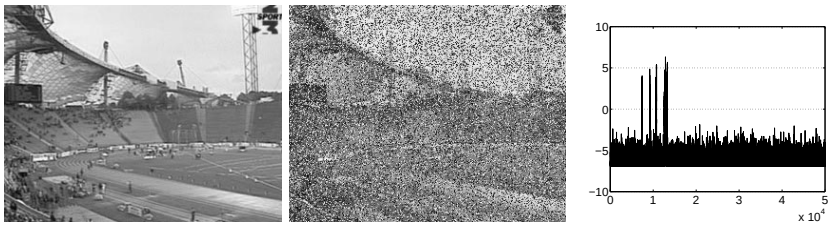

Fig. 2. The middle image is a 50% impulse noise version of the original one. In a database of 10^5 images, they still match with a NFA close to 10^{-5}. The right plot shows the confidence values $(-\log_{10}(NFA))$ for the first 50000 images of the sequence, the query being the noisy image. The peaks indeed correspond to exactly the same view of the stadium.

We first consider the following experiment. We select a single image in a sequence containing about one hour program of an athletics meeting (86096 images). This image represents a view of the stadium. To make the problem still more complex and to evaluate the robustness to noise, a white Gaussian noise with standard deviation $\sigma = 30$ is added to this image, and the resulting image will be taken as the query. The proposed criterion is applied with $M = 500$ random sample points in the images. The true image was detected with a NFA equal to 10^{-14}. About 20 images (belonging to the same static shot) are detected around the true image, which is of course correct as well. Moreover, this very same view of the stadium appears three other times in the video (before the selected true image). All of them are detected with a very low NFA (or equivalently, with a high confidence value, as shown in Fig.2). There was a single true false alarm (unrelated image) with a NFA equal to $10^{-0.73}$, which was probably due to the presence of the logo, but this NFA is coherent with the prediction: it is close to 1. No false alarms were obtained for an impulse noise of 50%. We have also applied JPEG compression to the original images. Extreme JPEG compression (quality less than 10) may lead to false detections since gradient orientation is constrained by the blocking effect. For usual compression ratio (quality 75), this effect was not observed.

On Fig. 3, two images of a movie are compared. The scene exhibits a strong transparency effect and an important contrast change. Thus, the grey levels in those images are different. Obviously, image intensity is not a good criterion at all, since the images apparences are different although the images clearly have a common cause. The gradient direction comparison proves that these images are similar in the sense that there resemblance cannot be explained by the *a contrario* model. It was empirically checked that sample points were quite uniformly distributed in the images. This experiment demonstrates that we are able to assess that two images are similar even if they are affected by transparency effects.

Fig. 3. Robustness to transparency. The two images are selected from a movie. The background is fixed, but the contrast changes a lot and a transparency layer is also moving. Nevertheless, with 200 sample points, the confidence value is $-\log_{10}(NFA) = 43.2$, and images are thus detected as very similar.

Fig. 4 shows the robustness to partial occlusion. The score panel occludes the bottom part of the image in this video of tennis match. The two images are detected as very similar since their NFA is about 10^{-50}. Since an hour of video contains about 10^5 images, such a NFA value asserts that the image pairing remains meaningful for any size of database. The threshold on the image gradient norm is equal to 5 in this experiment. If we take it equal to 0.2 (still with 200 sample points), the NFA increases since we select points where the gradient orientation is dominated by quantization. However, with an equal probability, we select points with larger gradients, and the gradient directions then match very well. Therefore, the NFA is still very low, and about 10^{-32}.

Fig. 4. Robustness to occlusion. Despite the partial occlusion the two images are detected as very similar with confidence value of $-\log_{10}(NFA) = 50.1$. The right plot gives the position of the 200 sample points. There are not points in constant areas (because of the gradient threshold). However, some points are selected in the non-matching area (scores), but the NFA is still very low.

As a last experiment, let us give a short insight of how geometrical invariance might be taken into account. We apply exactly the same decision scheme to pairs of consecutive images in a video sequence, but we first register the images by using the robust multiresolution motion estimation method by Odobez and Bouthemy [12], (the corresponding Motion-2D software is available on line at http://www.irisa.fr.vista/Motion2D) which computes a 2D parametric motion model that corresponds to the dominant image motion, which is usually related to the camera motion. The evolution of the NFA through time is represented on Fig. 5 (more precisely, the confidence values given by $-\log_{10}(NFA)$ are plotted). It indicates if the consecutive images of the video sequence (once registered) can be stated as similar or not. As expected, confidence is high in case of similarity since NFA are always lower than 10^{-20}, except at very precise instants that correspond to shot changes. Let us point out that an accurate registration of the two images to be compared is nevertheless required to properly exploit the proposed method for image similarity detection.

6 Conclusion and perspectives

We have described a novel and fast method allowing us to efficiently compare two images from a random sampling of points and to decide whether they are

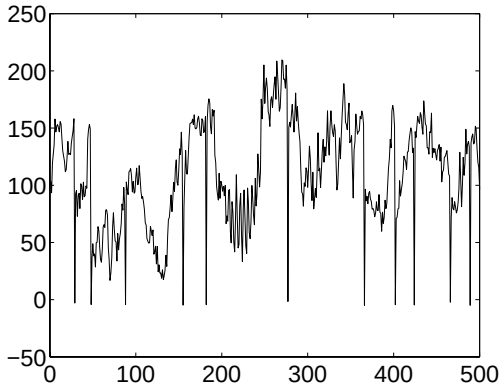

Fig. 5. Similarity evaluation between successive images of a video stream after registration. Plot of the confidence values $-\log_{10}(NFA))$ for 500 consecutive pairs in a MPEG video sequence. Most of the time, the NFA is below 10^{-20}. The sudden drops correspond to shot changes. The NFA is thus a reliable value as predicted by Proposition 1.

actually similar or not. It can be used for image comparison and image retrieval in databases or in video stream. Actually, the argument is quite general and the thresholds are rigorously proved to be robust and can be fixed once for all, for any type of images. Hence the user does not have to tune any parameter. Preliminary results have demonstrated the accuracy and the efficiency of the proposed method. Nevertheless, a more extensive experimental evaluation could be carried out. As an extension, our approach could also be applied to parts of images instead of entire images, so that the methodology could be used in many other applications of image retrieval, image matching or registration evaluation. These parts of images could be extracted from local characteristics as keypoints [11] or local frame based on stable directions [7,14]. We could then estimate the same detection bounds for system similar to [15]. This work is in progress.

A Proofs

Proof of Prop. 3. From (4), we know that $1-P_D = B(M, M-k, 1-p)$. A refined Stirling inequality [3] implies that

$$\begin{aligned}\frac{\varepsilon}{N_{\mathcal{B}}N_{\mathcal{Q}}} &> B(M, M-k, 1-p)\\ &\geqslant \binom{M}{M-k}(1-p)^{M-k}p^k\\ &\geqslant \frac{2}{\sqrt{2\pi M}}e^{-1/6}e^{-MH(1-k/M,1-p)}.\end{aligned}$$

Thus

$$M \cdot H\left(1-\frac{k}{M}, 1-p\right) > C + \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon \sqrt{M}},$$

with $C = \frac{1}{6} + \frac{1}{2} \ln \frac{\pi}{2} \simeq 0.39246$. Since $k > Mq$, we also have $H\left(1-\frac{k}{M}, 1-p\right) < H(1-q, 1-p)$. By convexity of H,

$$H(1-q, 1-p) \leqslant (p-q) \partial_x H(1-q, 1-p) = (p-q) \ln \left(\frac{1-q}{q} \frac{p}{1-p}\right).$$

Moreover

$$\ln \left(\frac{1-q}{q} \frac{p}{1-p}\right) = \int_q^p \frac{dx}{x(1-x)} \leqslant (p-q) \max_{x \in [p,q]} \frac{1}{x(1-x)}.$$

Since the function on the right hand side is convex, it attains its maximum on the boundary of the interval, and this completes the proof. □

Proof of Prop. 4. We first prove the following lemma, bounding from above the number of samples necessary to pass the test of similarity.

Lemma 1. *Let us fix $M > 0$ and $L = 1$ and let k be the minimal number of samples with similar directions such that the pair (u, v) is ε-meaningful.*

$$k \leqslant 1 + Mq + \left(\frac{M}{2} \left(\ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon}\right)\right)^{1/2}. \tag{8}$$

Proof. Since $k = \inf\{j \text{ s.t. } N_{\mathcal{B}} N_{\mathcal{Q}} \cdot B(M, k, q) < \varepsilon\}$, $B(M, k-1, q) > \frac{\varepsilon}{N_{\mathcal{B}} N_{\mathcal{Q}}}$ holds, also yielding

$$H\left(\frac{k-1}{M}, q\right) < \frac{1}{M} \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon}.$$

Convexity properties of the entropy H yield $H(r, q) \geqslant 2(r-q)^2$. Setting $r = \frac{k-1}{M}$ gives the result. □

If M is large enough, we can assume that $k < Mp$ from (8). A sufficient condition to $\mathcal{M}(M, P) < \varepsilon$ is

$$H\left(1-\frac{k}{M}, 1-p\right) > \frac{1}{M} \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon}$$

Since by convexity $H(r, p) \geqslant 2(r-p)^2$, it suffices that

$$2\left(p - \frac{k}{M}\right)^2 \geqslant \frac{1}{M} \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon},$$

which is implied by

$$p - q - \left(\frac{1}{2M} \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon}\right)^{1/2} > \left(\frac{1}{2M} \ln \frac{N_{\mathcal{B}} N_{\mathcal{Q}}}{\varepsilon}\right)^{1/2},$$

and the result directly follows. □

References

1. R. Brunelli, O. Mich and C.M. Modena. A survey on the automatic indexing of video data. *Jal of Visual Communication and Image Representation*, 10(2):78–112, 1999.
2. A. Desolneux, L. Moisan, and J.M. Morel. A grouping principle and four applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(4):508–513, April 2003.
3. W. Feller. *An Introduction to Probability Theory and its Applications*, volume I. J. Wiley, 3rd edition, 1968.
4. M.A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
5. W.E.L. Grimson and D.P. Huttenlocher. On the sensitivity of the Hough transform for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(3):255–274, 1990.
6. W. Hoeffding. Probability inequalities for sum of bounded random variables. *J. of the Am. Stat. Assoc.*, 58:13–30, 1963.
7. J.L. Lisani, L. Moisan, P. Monasse, and J.M. Morel. On the theory of planar shape. *SIAM Multiscale Mod. and Sim.*, 1(1):1–24, 2003.
8. J.L. Lisani and J.M. Morel. Detection of major changes in satellite images. In *IEEE Int. Conf. on Image Processing*, ICIP'03, Barcelona, Sept. 2003.
9. D. Lowe. Object recognition from local scale-invariant features. In *IEEE Int. Conf. on Computer Vision*, ICCV'99, Corfu, Sept. 1999.
10. D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Jal of Computer Vision*, 60(2):91–110, 2004.
11. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. Jal of Computer Vision*, 65(1-2):43 - 72, November 2005.
12. J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Visual Communication and Image Representation*, 6(4):348–365, 1995.
13. J. Peng, B. Yu, and D. Wang Images similarity detection based on directional gradient angular histogram. *16th Int. Conf. on Pattern Recognition*, ICPR'02, Quebec, August 2002.
14. C.A. Rothwell. *Object Recognition Through Invariant Indexing.* Oxford Science Publications, 1995.
15. J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *IEEE Int. Conf. on Computer Vision*, ICCV'03, Nice, Oct. 2003.
16. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
17. A. Venot, J.F. Lebruchec, and J.C. Roucayrol. A new class of similarity measures for robust image registration. *Computer Vision Graphics and Image Processing*, 28:176–184, 1982.
18. T. Veit, F. Cao and P. Bouthemy. Probabilistic parameter-free motion detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR'04, Washington D.C., June 2004.