Using extreme value theory for image detection

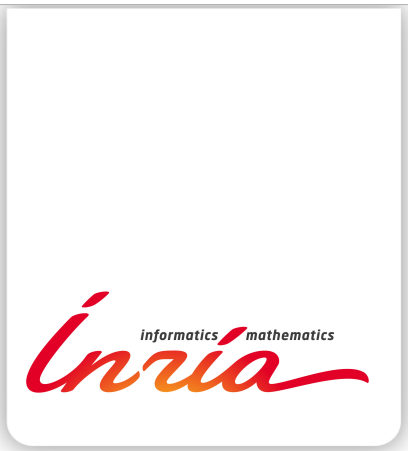
Teddy Furon, Hervé Jégou

Using extreme value theory for image detection

Teddy Furon, Hervé Jégou

RESEARCH REPORT
N° 8244
February 2013
Project-Team Texmex

ISSN 0249-6399 ISRN INRIA/RR--8244--FR+ENG



Using extreme value theory for image detection

Teddy Furon, Hervé Jégou

Project-Team Texmex

Research Report n° 8244 — February 2013 — 16 pages

Abstract: The primary target of content based image retrieval is to return a list of images that are most similar to a query image. This is usually done by ordering the images based on a similarity score. In most state-of-the-art systems, the magnitude of this score is very different from a query to another. This prevents us from making a proper decision about the correctness of the returned images.
This paper considers the applications where a confidence measurement is required, such as in copy detection or when a re-ranking stage is applied on a short-list such as geometrical verification. For this purpose, we formulate image search as an outlier detection problem, and propose a framework derived from extreme values theory. We translate the raw similarity score returned by the system into a relevance score related to the probability that a raw score deviates from the estimated model of scores of random images. The method produces a relevance score which is normalized in the sense that it is more consistent across queries.
Experiments performed on several popular image retrieval benchmarks and state-of-the-art image representations show the interest of our approach.

Key-words: image search, image detection, nearest neighbor, extreme value theory

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Détection d'images par valeurs extrêmes

Résumé : L'objectif de la recherche d'image est de retourner une liste d'images qui sont les plus visuellement similaires à une image requête, par exemple en ordonnant les images en fonction d'un score de similarité entre descripteurs d'images. Dans la plupart des systèmes de l'état de l'art, les scores varient significativement d'une requête à l'autre, ce qui empêche de déterminer la qualité intrinsèque des résultats retournés par la requête.

Cet article considère des applications pour lesquelles une telle mesure de confiance est requise, comme en détection de copies ou pour le reclassement d'image avec un système de vérification géométrique. Dans cet objectif, nous formalisons le problème de la recherche d'image comme un problème de détection d'*outlier*, en utilisant le cadre mathématique de la théorie des valeurs extrêmes. Nous transformons le score de similarité en une probabilité que le score dévie du modèle de score des images quelconques. Cette probabilité est un nouveau score de pertinence qui est comparable d'une requête à l'autre, et utilisé comme une mesure de confiance.

Des expériences effectuées sur des jeux d'évaluation usuels et plusieurs systèmes de recherche d'images montrent l'intérêt de notre approche.

Mots-clés : recherche d'image, détection d'image, plus proches voisins, théorie des valeurs extrêmes

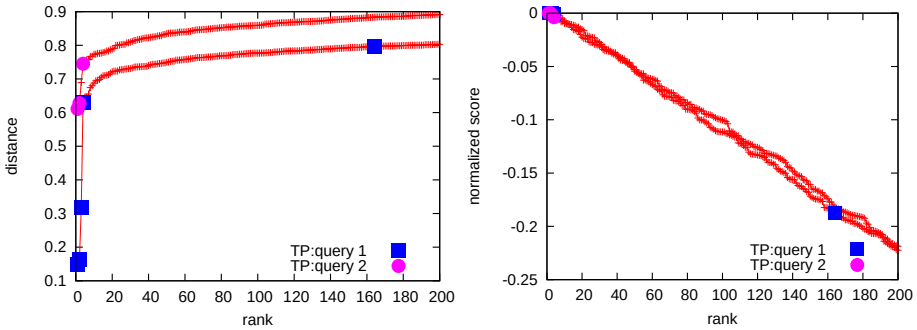


Figure 1: Illustration of the thresholding problem: *Left:* Two queries with the GIST descriptor in 1 million images. It is not possible to threshold the raw scores (here, Euclidean distances) to separate true positives from false positives. *Right:* The same queries, where our method have 'normalized' the raw scores into more meaningful relevance scores (log-likelihoods).

1 Introduction

Content Based Image retrieval (CBIR) is a historical line of research in Computer Vision which receives a continued attention from the community because of numerous applications. The problem usually consists in finding the images in a database that are most similar to a query. In recent years, many solutions have improved the search quality [1, 2, 3, 4, 5], while being more scalable. It enables to search in image sets comprising millions to thousand millions of images [6, 7]. Most of the proposed search techniques try to optimize the ranking, which typically corresponds to the "Google-image" application where a user has an interface presenting a fixed number of results.

In this paper, we are interested in a close problem: the *detection* of correct images. It shares a lot with the retrieval problem but yet has received less attention from the community [8]. The key difference between retrieval and detection is that detection requires the system to decide whether an image is correct or not. Thus it returns a set of results of variable size, since the number of relevant images depends on the query. The comparison of the scores output by the engine to a threshold decides which images are relevant, or even that no image is relevant. This is the typical way in the copy detection task of the evaluation campaign of TRECvid [9].

Another case of interest is when a post-verification, either manual or automatic (*e.g.*, a geometrical check [10, 3]), is performed. Thinking of it as a detection problem allows the optimization of the trade-off between the overall search quality and the complexity of the verification step. Section 2 describes other contexts where detection is of interest, along with some proper evaluation metrics for this problem.

As we will see, a good image retrieval system may be bad at taking decision. The scores output by the search engine are very dissimilar across queries [8]. Their comparison to an absolute threshold yields unreliable decisions as shown in Figure 1. This problem is related to the notion of meaningful nearest neighbors [11], *i.e.*, which was specifically raised in a context of matching. A few works [10, 12] proposed some rules to determine whether a local descriptor is

relevant or not. The popular Lowe's distance ratio criterion [10] discards a descriptor if the nearest neighbor is not significantly closer than the second one. The method of Omercevic et al. [12] weights the contribution of local descriptors when computing the similarity between two images. For a given query, they analyze the distribution of the nearest neighbor distances and fit an exponential distribution representing the background distribution. This determines to which extent the closest ones deviate from the other measurements. The assumptions proposed in this paper are mainly empirical ones, as well as the weighting strategy. Yet, it was shown to provide good results in practice for local SIFT descriptors.

The methodology proposed in our paper is related with the notion of meaningful nearest neighbors. We find in the extreme value theory [13] a rigorous mathematical framework supporting this concept. This concept was used in [14] for normalizing scores produced for several attributes by a support vector machine, thereby improving the fusion scheme. Our work departs from this work on the application, but also because we more specifically consider the problem of *outliers*: The scores of true positives should not be taken into account in the extreme value law estimation.

Our work also shows that the weighting scheme proposed by Omercevic et al. [12] is explained by extreme value theory, and that it has several limitations which are overcome by our framework. First, the exponential law assumption, which was empirically observed in this prior work, is theoretically explained. Yet, it is only one of the three possible cases that may occur when dealing with extreme values. Second and similar to [14], it does not consider the problem of outliers, which are important for a better parameter estimation [15]. Finally, the goal of our paper departs from these prior work [12, 14] by considering image detection and demonstrating its importance in this context. Our problem is therefore more related to the work of Perronnin et al. [8] from the application point of view.

We summarize our contributions as follows:

- We show the interest of a normalization scheme based on extreme value theory for image detection. Unlike related works in computer vision, but similar to works in other fields [15], it specifically takes into account the notion of outliers. It only requires to post-process the raw scores produced by an image engine at a negligible cost.
- We give some formal explanations to the empirical observations made by Omercevic et al. on the distribution of nearest local descriptors, and show the limitations of their approach.
- As a result, we report state-of-the-art performance on detection tasks.

Similar to former works using extreme values [15, 14], the technique is completely generic and can be applied on outputs of any search engine. This is in contrast, for instance, to the work of [8], which is specific to distribution based description and can therefore not be used with more complex (better) representations [16, 5].

2 Problem statement

This section presents some image search applications where producing meaningful scores, *i.e.*, consistent across queries, is of utmost importance. They are all characterized by the need of taking *a decision*, whence *a variable* number of results (possible none) for different queries. It also presents some evaluation metrics and protocols associated with these applications.

2.1 Detection of relevant images

The primary goal of image search is to return images from the database which are visually relevant to the query, for instance representing the same building [17], scene or object [2, 5]. The most common use-case considers that the user has an interface in which the top k images are displayed from the most to the less relevant. In this setup, the user satisfaction is reflected by the quality of the first results, and this is measured by the precision on the first results and/or the position of the correct images. In order to reflect a score independent of the choice of k, the typical measure in this context is the mean average precision (mAP), as implemented in [17]:

- The average precision (AP) is computed for each query ;
- These AP values are averaged over all queries.

Notice that this measure does not compare the scores across queries, and therefore it is not meant to capture the absolute quality of the results. It is not designed, in particular, to determine the number of images that are relevant, as shown by Figure 1. In other terms, this metric is not suitable for applications where the search engine takes decisions whether there is one, multiple, or even no matching image in the dataset.

In copy detection, a decision has to made whether the top-ranked candidates are copyright infringements of the queries. This has to be done based on the scores and therefore requires to fix a threshold once for *all* the queries. This is the game to be played in the copy detection task of TRECvid [9]. In this campaign, the evaluation metric is the normalized detection cost ratio (NDCR), which is a score between 0 and 1 (the lower, the better) of the form $\alpha\,\mathrm{FN}+\beta\,\mathrm{FP}$. The constants $\alpha > 0$ and $\beta > 0$ balance the cost of returning a false positive (FP) with that of missing a true positive, *i.e.* a false negative (FN).

2.2 Geometrical or manual verification

In order to scale to millions of images and yet to give precise results, the state-of-the-art systems [17, 5] adopt a two stage procedure where an efficient sorting first returns a short list of images deemed relevant, which might include FPs; and then a precise geometrical matching system filters this list. This spatial verification is costly and typically applied to check a few hundred images only. In some other applications such as media supervision, the search system is interactive and the user is asked to determine visually which images in the short list are relevant.

In this context, the objective is to obtain the best trade-off between the average size of the short list and the risk of missing a relevant image. The most

Dataset name	# images	# queries	# TP	# TP per query (*avg*)
Holidays [5]	1,491	500	991	1–12 (*1.98*)
Holidays+Flickr1M [5]	1,491 +1M	500	991	1–12 (*1.98*)
UKB [2]	10,200	10,200	10,200	4–4 (*4*)
Oxford105k [17]	5,062+100k	55	2840	6–221 (*51.6*)

Table 1: Datasets and images representations used in this paper.

common solution [5] is to return a fixed number of images before spatial or manual verification. It is not satisfactory because the queries have variable number of relevant images, and the results have considerably different reliabilities from a query to another.

2.3 Evaluation metrics for the detection problems

We adopt the choice of computing the AP jointly for all the queries, as suggested in [8]. This is done by sorting/interleaving the results of all queries based on their scores, and computing in turn the precision-recall curve globally. This "global" curve reflects to the aforementioned trade-off between the number of verifications to be performed and the rate of correct images that are short-listed.

To avoid any terminology ambiguity, this average precision is referred to as the global average precision (GAP) in the rest of the paper. This quantity synthetically aggregates into a single value the different possible trade-offs between the number of detected images by the system and the rate of detected relevant images. In this paper, it is used to compare the respective merits of image search engines with respect to image detection.

As a complementary way to compare the systems, we also consider the receiver operating curve (ROC) to compute the area under curve (AUC) synthetic measure. Although this curve is less popular in search applications, it is a standard way to evaluate the quality of a detector.

2.4 Datasets and image representations

Datasets. Table 1 summarizes the different datasets in this paper. They were originally introduced to gauge the quality of image search in a retrieval framework. In the University of Kentucky Benchmark, all the images are used as queries, while in Holidays only a subset is submitted to the search engine in a leave one out fashion. (it amounts to a database of 1490 images). For the Oxford105k building dataset [17], the queries are cropped images of some dataset images from the Oxford5k subset. Some "junk" images, whose relevance is ambiguous, are removed prior to evaluation. Please refer to the publications cited in the table for a full description of these datasets.

Image representations. The technique we are looking at is generic. It processes the short-list of scores returned by any kind of image search system. To demonstrate this experimentally, we consider four popular image representations and corresponding search engines, namely:

- The Bag-of-words (BOW) representation [1] with 20,000 centroids, which is obtained from local SIFT [10] computed on patches extracted by the Hessian-Affine detection [18] ;
- The improved BOW method based on Hamming Embedding and Weak Geometry Consistency [5], denoted by HE+WGC, which adds binary signatures and includes some partial geometrical information to refine the descriptor representation ;
- The global GIST description [19], which captures the global envelope of an image, and is particularly suited for scene recognition ;
- The recent improved Fisher Vector (FV) representation [20, 21] by Perronnin et al. It was especially shown to be successful in classification tasks, yet the authors also reported some competitive results in image retrieval [16].

For all these representations, we have used the code or descriptors shared online by different computer vision groups, or used the output of their search engines. We have not used any geometrical verification step, since our algorithm takes place before it and in order to optimize the trade-off between the size of the short list and the search quality.

The comparison metric is either a similarity or a dissimilarity, depending on the representation. In the following, we take the negation of all dissimilarities, so that, in all cases, the top-ranked images have the highest values. These data are the inputs of our algorithm, and are called the raw scores in the sequel.

3 A primer on extreme value theory

Extreme value theory has become over the last 50 years an essential branch of statistics. Inferring the statistical properties of super big but super rare observations is now the pivotal technique in risk management, portfolio adjustment, traffic, earthquake or flood prediction, etc. This section gives a brief overview of the main results of extreme value theory based on [13].

3.1 Main theorems of extreme value theory

We record a sequence of n scalar values denoted $x_1, x_2, \ldots, x_n$ that we model as continuous random variables, ie. X_i is assumed to have a probability distribution $F_i(x) = \mathbb{P}(X_i \leq x)$. Moreover, we assume that these observations are statistically independent and identically distributed (i.i.d.): $F_i(x) = F(x)$, $\forall i$.

The first pillar of extreme value theory deals with the maximum $M_n = \max(X_1, \ldots, X_n)$ for n going to infinity. The Fisher–Tippett–Gnedenko theorem establishes the asymptotical probability distribution of M_n (see [13, Th. 3.1]):

Theorem: *If there exist a sequence of normalizing constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\mathbb{P}(a_n^{-1}(M_n - b_n) \leq z) \to G(z) \quad \text{as } n \to \infty, \tag{1}$$

then G *belongs to one of the three following family: Gumble, Fréchet, and Weibull.*

Indeed, the three families can be expressed by a single formula, so-called generalized extreme value (GEV) distribution:

$$G(z;\mu,\sigma,\xi) = \exp\left(-\left(1+\xi\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right), \tag{2}$$

where μ is the location parameter, $\sigma > 0$ is the scale parameter, and ξ is the shape parameter. Fréchet and Weibull corresponds to $\xi > 0$ and $\xi < 0$ respectively. The Gumbel distribution is the limit when $\xi \to 0$: $G(z;\mu,\sigma,0) = \exp(-\exp(-(z-\mu)\sigma^{-1}))$.

The remarkable fact is the universality of the theorem: there are only three extreme value distributions regardless of the distribution of the observations $F(x)$. This theorem is often considered as an analog of the central limit theorem for the max operator. However, it doesn't fit well in the context of CBIR. If we were observing several maxima, each of them taken from an independent set, then we can think about modeling the distribution of these maxima by a GEV. In CBIR, we have a unique database of n objects (images or descriptors), and the system returns the k biggest similarity measures. This is the reason why we resort to the Pickands–Balkema–de Haan theorem (see [13, Th. 4.1]):

Theorem: *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with distribution $F(x)$ such that Theorem 1 holds. Then, for a large enough threshold u, the distribution of the threshold excesses, ie. the variables $(X_i - u)$ if they are positive, has a limit given by:*

$$H(y;\xi,\underline{\sigma}) = \mathbb{P}(X-u \le y | X > u) = 1 - (1+\xi y/\underline{\sigma})^{-1/\xi} \tag{3}$$

with $\underline{\sigma} = \sigma + \xi(u-\mu)$. This distribution is defined for $y > 0$ and $(1+\xi y/\underline{\sigma}) > 0$.
This family of distribution is called the Generalized Pareto (GP) distributions. The parameter ξ is of utmost importance because it qualifies its shape:

- $\xi < 0$ implies that the threshold excesses are upper bounded by $-\underline{\sigma}/\xi$,
- $\xi > 0$ implies that the threshold excesses are not bounded and decaying with an heavy tail,
- $\xi = 0$ implies that the threshold excesses are not bounded and decaying exponentially:

$$H(y;0,\underline{\sigma}) = 1 - \exp\left(-\underline{\sigma}^{-1}y\right), \quad y > 0.$$

Once again, if $F(x)$ is known, then it is easy to get parameters $(\mu,\underline{\sigma})$. Here are two examples:

Example 3.1: For the exponential model $E(\lambda)$, $F(x) = 1-\exp(-\lambda x)$ for $x > 0$, we have:

$$H(y;\xi,\underline{\sigma}) = 1 - \mathbb{P}(X-u > y | X > u) = 1 - \frac{1-F(u+y)}{1-F(u)} = 1 - \exp(-\lambda y),$$

which is still an exponential, ie. a GP distribution with $\xi = 0$, $\underline{\sigma} = \lambda^{-1}$.

Example 3.2: For the uniform model $U(0,1)$, $F(x) = x$ for $0 \leq x \leq 1$, we have:

$$H(y;\xi,\underline{\sigma}) = 1 - \mathbb{P}(X - u > y | X > u) = 1 - \frac{1-(u+y)}{1-u} = \frac{y}{1-u},$$

which is still a uniform distribution $U(u,1)$, ie. a GP distribution with $\xi = -1$, $\underline{\sigma} = 1 - u$.

Note that these two examples are exceptions where the property hold whatever the threshold $u > 0$, *i.e.*, it doesn't need to be large. Yet, if $F(x)$ is unknown but assumed to satisfy Theorem 1 (and, as far as we know, all 'textbook' distributions with finite variance do), then we have the distribution of the threshold excesses provided u is large enough and if we can estimate $(\xi, \bar{\sigma})$.

3.2 Estimation of the GP distribution parameters

We observe a sequence of values $x_1, x_2, \ldots, x_n$ and we assume that the data have been already sorted in decreasing order: $x_1 > x_2 > \ldots > x_n$. We set the value of the threshold as $u = x_k$, and compute the threshold excesses as $y_i = x_i - u$ for $1 \leq i \leq k$. This imposes that $k \ll n$ so that u is large enough in order to apply theorem 2. In the other hand, k must not be too small as we estimate the distribution parameters from $\{y_i\}_{i=1}^k$. As a rule of thumb, k should be some tenths so that n must be some thousands.

We use the Maximum Likelihood Estimator (MLE), which amounts to find the maximizers $(\hat{\xi}, \hat{\underline{\sigma}})$ of the following function:

$$\ell(\xi,\sigma) = -k\log\sigma - (1+1/\xi)\sum_{i=1}^{k}\log(1+\xi y_i/\sigma). \tag{4}$$

This needs some precautions to avoid numerical instabilities, in particular when $\xi \approx 0$. We indeed set a limit $\xi_{\lim} > 0$. The MLE is first run on the domain $\xi_{\lim} \leq \xi \leq 1$ (the mean of GP variable is infinite for $\xi > 1$) and $0 \leq \sigma$ to give $(\hat{\xi}^+, \hat{\underline{\sigma}}^+)$. We run again the MLE on the domain $-1 \leq \xi \leq -\xi_{\lim}$ (MLE is known to be unstable for $\xi < -1$) and $0 \leq \sigma$ to give $(\hat{\xi}^-, \hat{\underline{\sigma}}^-)$. In this domain, $\ell(\xi,\sigma) = -\infty$ if $\exists y_i : (1 + \xi y_i/\sigma) \leq 0$. We now consider the case where $\xi = 0$ and the likelihood is:

$$\ell(0,\sigma) = -k\log\sigma - \sigma^{-1}\sum_{i=1}^{k} y_i. \tag{5}$$

The maximizer is denoted $\hat{\underline{\sigma}}^{(0)}$. The final estimates are set depending on the maximum of the three local maxima $\ell(\hat{\xi}^-, \hat{\underline{\sigma}}^-)$, $\ell(\hat{\xi}^+, \hat{\underline{\sigma}}^+)$, and $\ell(0, \hat{\underline{\sigma}}^{(0)})$.

To ease the maximization, we can parametrize σ as a function of ξ. For instance, a GP variable Y satisfies $\mathbb{E}_Y[y] = \sigma/(1-\xi)$ provided $\xi < 1$. Therefore, we transform the maximization over two variables (ξ, σ) by a maximization of the scalar function $L(\xi) = \ell(\xi, (1-\xi)k^{-1}\sum_{i=1}^{k} y_i)$. Other re-parametrization method are suggested in [13] under the name of profile log-likelihood.

4 Scoring by modeling extreme values

For a given query image, the CBIR system produces the raw scores of the top-k images in the database: $x_1 \geq x_2 \geq ... \geq x_k$. The above section explained that we can have a statistical model of these data even if we ignore the distribution of X_i. The only assumption is that the scores are occurrences of random variables which are i.i.d, *i.e.*, distributed according to an unknown but unique law. This is obviously not true in practice in CBIR because the raw scores are a mix of:

- scores of relevant images, which do share a similarity with the query,
- scores of irrelevant images. These scores are so big just because they are the extreme values over a big set of irrelevant image scores that we assume to be i.i.d.

This motivates the second part of our algorithm: the detection of outliers.

4.1 Detection of outliers

We restrict the detection to the following case: we assume that, if there are some outliers, they have the biggest scores. Suppose for the moment, there may be one outlier. The question is whether the biggest score x_1 is distributed according to the GP distribution whose parameters have been estimated with $\{y_i\}_{i=1}^k$. Our solution is to consider x_1 has an empirical quantile of the distribution $F(x)$. This score is the biggest over a set of size n. Therefore, it should be close to the theoretical quantile $F^{-1}(1-1/n)$. The following theorem gives a strong support to this rationale (see [22, Th. 7.25]).

Theorem: *Suppose $X_1 \geq X_2 \geq \ldots \geq X_n$. Denote by $Q_{p,n} = X_{n-[np]}$ the empirical quantile of order p. For all $p \in (0,1)$, if $F(x)$ is continuous in p and its quantile of order p equal to $x^{(p)}$ (ie. $F(x^{(p)}) = p$), then*

$$Q_{p,n} \to x^{(p)}, \quad \textit{almost surely as } n \to \infty. \tag{6}$$

Moreover, if $F(x)$ is derivable in a probability density function $f(x)$ which is strictly positive in the neighborhood of $x^{(p)}$, then

$$\sqrt{n}(Q_{p,n} - x^{(p)}) \to \mathcal{N}(0, \sigma_p^2), \quad \textit{almost surely as } n \to \infty, \tag{7}$$

with $\sigma_p^2 = p(1-p)/f(x^{(p)})^2$.

We use this theorem with $p = 1 - 1/n$ so that $Q_{p,n} = x_1$. We need the expressions of $x^{(p)} = F^{-1}(p)$ and $f(x^{(p)})$. Yet, we have been claiming from the beginning that $F(x)$ is unknown. The trick is that we only need its expression around $x^{(p)}$ which is bigger than u. We write the following expression:

$$\begin{aligned} F(x^{(p)}) &= \mathbb{P}(X \leq x^{(p)}) = \mathbb{P}(X \leq x^{(p)}, X > u) + \mathbb{P}(X \leq x^{(p)}, X \leq u) \\ &= \mathbb{P}(X \leq x^{(p)} | X > u)\mathbb{P}(X > u) + \mathbb{P}(X \leq u) \\ &\approx H(x^{(p)} - u; \hat{\xi}, \underline{\hat{\sigma}}).k/n + (1 - k/n) \end{aligned} \tag{8}$$

The conditional probability is replaced by the GP distribution with estimated parameters, while $\mathbb{P}(X > u) \approx k/n$ is the probability of being short-listed. This

gives the expression of the theoretical quantile:

$$x^{(p)} \approx \begin{cases} u + \hat{\underline{\sigma}}((k/n(1-p))^{\hat{\xi}} - 1)/\hat{\xi} & \text{if} \quad \hat{\xi} \neq 0 \\ u + \hat{\underline{\sigma}} \log(k/n(1-p)) & \text{if} \quad \hat{\xi} = 0 \end{cases} \tag{9}$$

but also the expression of the pdf in $x^{(p)}$:

$$f(x^{(p)}) \approx \begin{cases} \left(1 + \hat{\xi} x^{(p)}/\hat{\underline{\sigma}}\right)^{-(1+1/\hat{\xi})} /\hat{\underline{\sigma}} & \text{if} \quad \hat{\xi} \neq 0 \\ \exp(-x^{(p)}/\hat{\underline{\sigma}})/\hat{\underline{\sigma}} & \text{if} \quad \hat{\xi} = 0 \end{cases} \tag{10}$$

Now, thanks to the second part of the theorem 3, we decide that x_1 is an outlier ($d = 1$) or not ($d = 0$) by

$$d = (\Phi((x_1 - x^{(p)})/\sigma_p) > 1 - \alpha), \tag{11}$$

where $\Phi(x)$ is the Gaussian distribution and α is the probability of false positive.

4.2 Our algorithm

Our algorithm consists in alternating the estimation of the parameters and the detection of one outlier. Although some works in computer vision have proposed to use extreme value for score normalization [14], it is more related to the work by Olmo [15] since it explicitly aims at removing the undesirable impact of outliers on the estimation. We first set $u = x_k$ and $y_i = x_i - u$ for $1 \leq i \leq k$. Denote this set $\mathcal{Y} = \{y_i\}_{i=1}^k$. The number of detected outlier n_o is initialized to 0. The detection output d is set to 1.

1. `while` $d = 1$,
   - $(\hat{\xi}, \hat{\underline{\sigma}})$ is estimated from $\mathcal{Y}$ by (4),
   - The detection is given by (11) based on $(\hat{\xi}, \hat{\underline{\sigma}})$,
   - if $d = 1$, then $n_o := n_o + 1$, $\mathcal{Y} := \mathcal{Y} \backslash \{y_{n_o}\}$,
2. Output: $s_i = H(y_k; \hat{\xi}, \hat{\underline{\sigma}})$ (or $t_i = \log(s_i/(1 - s_i))$) along with n_o

The final outcomes $\{s_i\}_{i=1}^k$ lie in between 0 and 1. Note that function $H(\cdot; \hat{\xi}, \hat{\underline{\sigma}})$ is a cumulative distribution function, hence an increasing function, so that this process doesn't change the order of the candidates for a given query. However, from a query to another, $(\hat{\xi}, \hat{\underline{\sigma}})$ are likely to differ and this produces a re-ranking. To increase the dynamic, we also use $\{t_i\}_{i=1}^k$ with $-\infty < t_i < +\infty$, and $\log(x/(1-x))$ is also an increasing function. As a side product, the algorithm also returns n_o in the sense that the top-n_o are deemed outliers.

5 Experiments

This section describes some experiments performed on various image retrieval benchmarks and description techniques. Please refer to Section 2.3 for a brief description. In contrast with most works of the literature which are focused on the retrieval performance, we evaluate the detection performance, which is measured either by the GAP (from the global average precision curve) or the

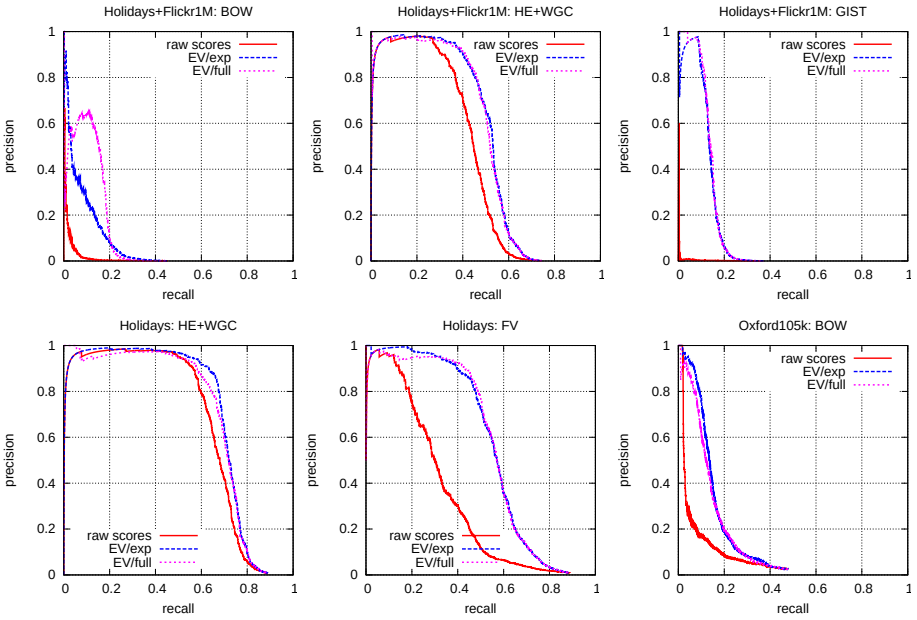


Figure 2: Global precision/recall curves on different datasets.

AUC measure (from the ROC curve), as explained in Section 2.3. We evaluate the two possible variants of the algorithm described in Section 4.2: the general method, denoted by EV/full, which estimates $\hat{\xi}$ over $(-1, 1)$, and the restricted variant, referred to as EV/exp, which assumes that the GP distribution is indeed an exponential distribution, ie. $\hat{\xi}$ is forced to 0.

Detection performance. Figure 2 shows that our method is effective and general. It is easily plugged on any existing system to improve the performance with respect to detection, without affecting in any manner the retrieval ability. Once gain, for a given query our method does not modify the ranking.

Table 2 gives a more synthetic view on more combinations of datasets and descriptors. Overall, the detection performance is in most of the cases significantly increased. The only counter-example is the Oxford105k with HE+WGC, where the model EV/exp gives the same performance as the raw scores, while the general model is not good.

Impact of database size. By comparing the relative detection gain, we conclude that our method is comparatively better with larger datasets. This is not surprising, because larger sets mean that the assumptions involved in extreme values theory are better satisfied.

Image representations. The improvement for the GIST descriptor is comparatively better. Our interpretation is that the GIST is a global descriptor which particularly suffers from the non stationarity of the feature space. In contrast, the other representations are derived from local descriptors, which occur to be more reliable for detection without any score normalization.

Comparison with the state-of-the-art detection systems. Table 3 shows

Benchmark	Description	GAP raw	GAP EV/exp	GAP EV/full	AUC raw	AUC EV/exp	AUC EV/full
Holidays	BOW	21.2	33.6	38.8	78.2	85.4	86.3
	HE+WGC	73.8	78.9	77.4	94.9	95.8	95.8
	GIST	11.8	36.5	36.6	72.1	83.3	83.4
Holidays+Flickr1M	BOW	2.5	15.3	23.5	72.9	87.1	79.3
	HE+WGC	57.9	68.0	67.5	93.1	95.3	95.3
	GIST	0.8	36.5	38.2	70.0	88.0	88.1
Oxford105k	BOW	14.3	32.5	29.9	73.3	77.7	76.8
	HE+WGC	65.0	64.3	51.5	88.5	88.2	87.1
UKB	BOW	51.2	76.2	75.1	89.6	96.2	95.7
	HE+WGC	86.6	91.0	89.0	97.3	98.0	98.0
	GIST	54.0	59.4	58.9	83.1	90.0	90.1

Table 2: Impact of the extreme value post-processing on different benchmarks and image search techniques. The AUC is computed from the ROC curve.

Contextual dissimilarities [8]

	mAP	GAP raw	GAP Ctxt
BOW/L2	45.7	18.5	37.9
BOW/L1	55.0	16.8	47.0

Proposed Extreme value method

	mAP	GAP raw	GAP EV/exp	GAP EV/full
BOW/L2	46.9	21.2	33.6	**38.8**
HE+WGC	79.4	73.8	**78.9**	77.4
improved FV [20, 21]	62.5	57.9	**68.0**	67.5
GIST	36.5	11.8	36.5	**36.6**

Table 3: Comparison with the state of the art on image detection: GAP measurement on Holidays. For reference, we provide the regular mAP as a gauge of retrieval performance. Our method does not change the ranking per query and therefore the mAP remains the same. *Legend: EV/exp forces* $\hat{\xi} = 0$*, while EV/full automatically finds the best* $\hat{\xi} \in (-1, 1)$*, see Section 4. The 'raw' columns give the GAP at the output of the search systems.*

the performance of different state-of-the-art systems with respect to the detection problem measured by the GAP. This done on Holidays like in the prior work of Perronnin et al. [8]. Our results for the BOW baseline are slightly better than theirs, yet remain comparable for the same L2 metric.

Note that the method of [8] only applies to distribution based description. It can not be used in conjunction with better image representations. As a result, although the contextual dissimilarities give a strong improvement on top of BOW, it is already outperformed by the recent indexing techniques, as shown by the GAP of raw scores for the HE+WGC and FV representations.

In contrast, our technique is generic and can be applied to any search engine. As a result, we significantly outperform the prior work by applying our technique jointly with better systems. On the BOW representation associated with the Euclidean distance, our results are comparable with those of Perronnin et al.

Complexity. Our method is a post-processing stage which estimates only two parameters from the list of the raw scores returned by the initial image search system. With a non-optimized Matlab implementation[1], processing all

[1]The Matlab package that reproduces some results of this paper will be shared along with

the queries for the datasets takes a few seconds, regardless of the database size. It is therefore negligible compared with the cost of the image retrieval system that returns the short list.

6 Discussion

Let us first summarize the features of our algorithm. It translates the short list of the top raw scores given by the search engine into a list of probabilities of being outlier. This monotonic mapping does not modify the ranking of the candidates in the short list. Therefore, it does not affect the performance for measure purely based on the rank like mean average precision. Yet, it is adaptive from a query to another, producing scores which are more homogeneous, and therefore more suitable for making a decision.

This translation is based on an extreme value distribution. When $\hat{\xi} = 0$, the score distribution is exponential (see Example 3.1 in Section 3), and our algorithm maps x_i into $s_i = 1 - \exp(-(x_i - x_k)\underline{\hat{\sigma}}^{-1})$. This is exactly the weighting proposed by Omercevic et al. [12] for SIFT descriptors. Yet, in our work, i) $\underline{\hat{\sigma}}$ is given by the MLE, a provably good estimator, ii) detected outliers are removed from the short list, iii) this is only one distribution among the GP family.

We must mitigate this criticism. It appears that $\hat{\xi} \approx 0$ in most of our experiments, especially when dealing with search engines whose similarity measure is a scalar product between two descriptors. For Holidays+Flickr1M, 80% of times $0 \leq \hat{\xi} < 0.2$ for HE+WGC. We explain this as follows: For a given query, the scalar product is a linear combination of the components of the candidate descriptor. If this vector is long enough, the raw score tends to be Gaussian distributed thanks to the central limit theorem, and the threshold excesses are exponentially distributed. This would be the case also if the search engine works with normalized vectors and Euclidean distance since $x_i = -\|\mathbf{d}_i - \mathbf{q}\|^2 = 2\mathbf{d}_i^\top \mathbf{q} - 2$. Yet, this does not apply with another distance or when the descriptors are not normalized vectors. For a general distance, say $x_i = -D(\mathbf{d}_i, \mathbf{q})$, raw scores are upper bounded by 0. Hence, a GP distribution with $\xi < 0$ is more likely to fit. In contrast with previous example, 90% of times $\hat{\xi} < -0.8$ for BOW on Holidays+Flickr1M. Note that in the limit $\xi = -1$, the mapping is just a rescaling of the raw score (see Example 3.2). These handwaving considerations show that, with some prior about the search engine, it might be possible to restrict the search of $\hat{\xi}$ to one of the three domains.

Last but not least, the main assumptions are i) $k \ll n$ and ii) k minus the number of outliers is large enough to enable an accurate parameter estimation. This explains why there is no real improvement for Oxford105k with HE+WGC: The number of TP is more than 50 on average, which strongly disturbs the MLE. In contrast, on the same dataset but using the BOW representation, the short list contains more false positive and the improvement is, ironically, much better.

the lists of raw scores.

7 Conclusion

This report has shown the interest of using extreme value estimation with outliers detection in the context of image detection. The algorithm produces a probabilistic confidence score per image. We believe that an important application of this technique is to estimate the number of results which are likely to be relevant to the query, *i.e.*, the short-list size. As a particular case, it allows the image search system to automatically determine how many images should undergo a spatial verification that will filter out the remaining outliers.

Acknowledgments: This project was partly done in the context of the Quaero Project and of the ANR Project Secular (ANR-12-CORD-0014).

References

[1] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003)

[2] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006) 2161–2168

[3] Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: ICCV. (2007)

[4] Perdoch, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: CVPR. (2009)

[5] Jegou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. IJCV **87** (2010) 316–336

[6] Torresani, L., Szummer, M., Fitzgibbon, A.: Learning query-dependent prefilters for scalable image retrieval. In: CVPR. (2009)

[7] Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010)

[8] Perronnin, F., Yan, L., Renders, J.M.: A family of contextual measures of similarity between distributions with application to image retrieval. In: CVPR. (2009)

[9] Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: MIR. (2006) 321–330

[10] Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110

[11] Baumberg, A.: Reliable feature matching across widely separated views. In: CVPR. (2000)

[12] Omercevic, D., Drbohlav, O., Leonardis, A.: High-dimensional feature matching: employing the concept of meaningful nearest neighbors. In: ICCV. (2007)

[13] Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer (2001)

[14] Scheirer, W.J., Kumar, N., Belhumeur, P.N., Boult, T.E.: Multi-attribute spaces: Calibration for attribute fusion and similarity search. In: CVPR. (2012)

[15] Olmo, J.: Extreme value theory filtering techniques for outlier detection. Technical Report 09/09, City University London - Department of Economics (2009)

[16] Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: CVPR. (2010)

[17] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007)

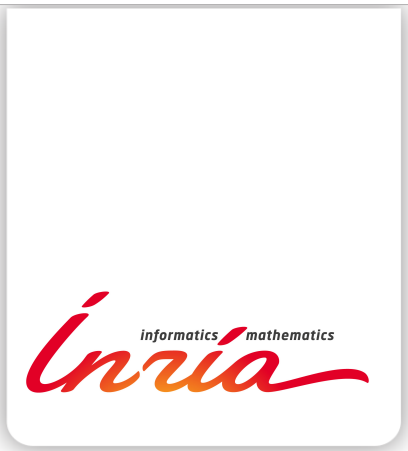
[18] Mikolajczyk, K., schmid, C.: Scale and affine invariant interest point detectors. IJCV **60** (2004) 63–86

[19] Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV **42** (2001) 145–175

[20] Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)

[21] Perronnin, F., J.Sanchez, Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010)

[22] Schervish, M.J.: Theory of statistics. Springer Series in Statistics. Springer-Verlag, New York (1995)



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399