On accuracy, robustness and security of bag-of-word search systems

Sviatoslav Voloshynovskiy*, Maurits Diephuis, Dimche Kostadinov, Farzad Farhadzadeh and Taras Holotyak
University of Geneva, Department of Computer Science, Stochastic Information Processing Group
7 route de Drize, CH 1227, Geneva, Switzerland

ABSTRACT

In this paper, we present a statistical framework for the analysis of the performance of Bag-of-Words (BOW) systems. The paper aims at establishing a better understanding of the impact of different elements of BOW systems such as the robustness of descriptors, accuracy of assignment, descriptor compression and pooling and finally decision making. We also study the impact of geometrical information on the BOW system performance and compare the results with different pooling strategies. The proposed framework can also be of interest for a security and privacy analysis of BOW systems. The experimental results on real images and descriptors confirm our theoretical findings.

Notation: We use capital letters to denote scalar random variables X and $\mathbf{X}$ to denote vector random variables, corresponding small letters x and $\mathbf{x}$ to denote the realisations of scalar and vector random variables, respectively. We use $\mathbf{X} \sim p_{\mathbf{X}}(\mathbf{x})$ or simply $\mathbf{X} \sim p(\mathbf{x})$ to indicate that a random variable $\mathbf{X}$ is distributed according to $p_{\mathbf{X}}(\mathbf{x})$. $\mathcal{N}(\mu, \sigma_X^2)$ stands for the Gaussian distribution with mean μ and variance σ_X^2. $\mathcal{B}(L, P_b)$ denotes the binomial distribution with sequence length L and probability of success P_b. $||.||$ denotes the Euclidean vector norm and $Q(.)$ stands for the Q-function. $\mathcal{D}(.||.)$ denotes the divergence and $\mathbb{E}\{.\}$ denotes the expectation.

1. INTRODUCTION

The BOW framework has been widely used in content search systems, biometric applications such as face or gait recognition and more recently in multimedia security applications including copy detection, black list tracking, content blocking and commercial content ranking systems. Modern BOW based systems can easily handle large-scale search or recognition problems, even on mobile phones. The BOW approach is based on the construction of a *visual alphabet* or *dictionary* based on the clustering of low-level features such as discriminative and robust descriptors.

Traditionally, *vector quantization* (VQ) is used for clustering based for example on k-means or another unsupervised learning algorithm.[7,9] The resulting visual alphabet represents a clustered structure with the centroids, or so called *visual words*, matching to the low level descriptors. In retrieval and classification applications, a training set of images is used to generate a set of features used in the above mentioned clustering. Once the visual alphabet is learned, all individual images in the database are represented in the form of a histogram of clusters, which can be efficiently indexed using an inverted file. A query image is also converted into the BOW space and the search system returns the database entries that have statistical properties close to those of the query. In this case, a global similarity measure between the images is approximated by a set of local similarity measures in the space of the learned visual alphabet.

Therefore, the crucial elements of BOW systems include: (a) selection or learning of robust, discriminative and invariant local/global descriptors; (b) construction of a visual alphabet and (c) efficient representation of enrolled images and the query in the BOW space by trading-off the robustness to distortions, distinguishability and memory storage which includes the so-called encoding and construction of a visual tree, (d) pooling the results to geometrically robust fixed-length vectors and finally (e) decision making.

The state-of-the-art in BOW research covers several main directions:

*The contact author is S. Voloshynovskiy (email: svolos@unige.ch). http://sip.unige.ch

- *descriptors*: the design of new very short (about 32 bits) binary descriptors such as CHOG[7] and ORB[15] are of great interest for mobile on-line applications in contrast to non-binary, long and computationally heavier SIFT[12] and SURF[1] descriptors;

- *encoding/assignment*: the design of new encoding/assignment strategies capable of providing an efficient representation of the descriptors in the codebook space, represented by visual words next to ensuring fast ϵ-NN or k-NN search. These techniques are closely related the various methods used to create accurate vector approximations of descriptors. The latter can be roughly classified into three groups:

  - *VQ (hard)-assignment*[16]: the ith descriptor $\mathbf{x}^i$ is quantized by a coarse vector quantizer $\mathbb{Q}_c(.)$ to its nearest visual word resulting in the approximate $\hat{\mathbf{x}}^i = \mathbb{Q}_c(\mathbf{x}^i)$;
  - *source coding with refinement*[9]: the descriptor $\mathbf{x}^i$ is quantized using the VQ $\mathbb{Q}_c(.)$ to its nearest visual word with simultaneous storage of the quantized refinement information of the descriptor with respect to its quantized version given by the fine quantization function $\mathbb{Q}_f(.)$ of the residual vector: $\hat{\mathbf{x}}^i = \mathbb{Q}_c(\mathbf{x}^i) + \mathbb{Q}_f(\mathbf{x}^i - \mathbb{Q}_c(\mathbf{x}^i))$;
  - *soft-assignment*[11, 19]: the descriptor $\mathbf{x}^i$ is approximated in the space of visual words $\mathbf{c}^j$, $1 \leq j \leq J$, by the linear approximation: $\hat{\mathbf{x}}^i = \sum_{j=1}^{J} w_{i,j}\mathbf{c}^j$. Strategies such as *sparse coding* and *locally-constrained linear coding* (LLC) include different selection strategies of their weight coefficients $\{w_{i,j}\}$;

- *pooling*[11]: the design of methods capable in dealing with a different number of features between the enrolled images and those in the probe, as well as ameliorating the lack of geometrical consistency between features. These methods are generally known as *sum/average* and *max pooling* methods;

- *analysis of security*[6, 13]: the investigation of different strategies to cope with descriptor attack sensitivity and all related methods that aim at protecting and enhancing the privacy and security of images enrolled in BOW based systems.

Nowadays, the design of existing BOW-based systems is based on memory/complexity considerations in view of the large-scale nature of the search problem. It includes a lot of heuristics and engineering, where performance is mostly evaluated by testing on (large) databases, and empirically compared for different descriptor classes and encoding algorithms. To the best of our knowledge there is no published work that rigorously links the robustness of the descriptors to overall system performance and its (optimised) parameters. It is still unclear what the gap is between non-synchronised systems based on pooling and synchronised systems based on geometric consistency validation. In fact the superiority of max pooling over the sum/average-pooling has only been shown in but a few experimental works. The impact of the accuracy of the feature representation or compression deployed in the visual codebook is mostly studied in an experimental way using an available dataset. In addition, the security of BOW-based systems is not completely explored and there is a lack of systematic study.

Therefore, in this paper, we are not interested in following some particular design of BOW-based systems but will rather consider the fundamental underlying assumptions, indicate the shortcomings and weaknesses of existing solutions next to proposing some solutions to these problems. On the other hand, we intend to introduce the strict apparatus of detection theory to evaluate the performance and security of BOW-based systems.

Currently, most BOW systems are used for CBIR, object recognition and copy detection. We will consider *content identification* where M items are enrolled and given a probe, the system should determine the corresponding item or issue a rejection. When it is not possible to return a single index item, the system should retrieve a list of indices whilst ensuring that the true item index is on the list. The content identification problem is traditionally analysed in the content fingerprinting formulation. The existing theoretical works[4, 14, 17, 18] mostly consider content identification based on one single sufficiently long fingerprint deduced to represent the content. In most theoretical works, perfect synchronisation between the enrolled fingerprint and the probe fingerprint is assumed with one notable exception[14] where the fingerprint de-synchronization was modelled by a random shift parameter. However, in practice it is difficult to design one single super fingerprint or descriptor that would be invariant to all types of distortions, which is why multiple short fixed-length local descriptors possessing certain invariance to geometrical transforms and robustness to signal processing operations are mostly used. However,

in this case the length of the deployed descriptors does not satisfy the asymptotic assumptions considered in the theoretical works[4, 14, 17, 18] which makes the analysis of practical BOW-systems intractable.

To the best of our knowledge there is little work on the theoretical analysis of BOW-systems' performance besides[21] and none on BOW based content identification. Therefore, the goal of this paper is to provide a simple and tractable model that allows to analyze, optimize and guide the design of BOW systems. In this paper, we will consider two cases of non-compressed and compressed features to reveal the theoretical limits of BOW based identification systems, analyze the impact of descriptor compression and encoding/assignment as well as discovering the impact of geometrical consistency between the descriptors on overall system performance. Such a formulation was not considered in earlier studies.

The BOW based content identification system is designed according to Figure 1. In this paper, we will not consider the geometric verification stage. Instead we are interested in establishing the list of possible candidates obtained by the feature matching step after different pooling strategies under the assumption that the features are not synchronised. To establish the upper performance limit of feature matching under the best achievable geometrical matching strategy, we also consider the case when the features are perfectly synchronised.

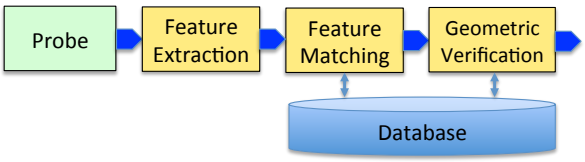


Figure 1. Block-diagram of BOW identification system.

The paper is organized as follows. The BOW based content identification problem is formulated in Section 2. Section 3 introduces the statistical model. Section 4 considers the performance of BOW content identification systems for the uncompressed features and Section 5 reveals the impact of feature compression. Section 6 introduces the security of BOW systems under informed attacks. Finally, Section 7 and Section 8 present the results and conclusions.

2. BOW-BASED CONTENT IDENTIFICATION: PROBLEM STATEMENT

A content database is defined as a collection of M items $\mathbf{x}(m)$ represented by their features/descriptors $\mathbf{x}(m) = \{\mathbf{x}^1(m), \cdots, \mathbf{x}^{J_{\mathrm{x}}(m)}(m)\}$, $1 \leq m \leq M$, with each descriptor $\mathbf{x}^i(m) \in \mathcal{X}^L$, $1 \leq i \leq J_{\mathrm{x}}(m)$ and $J_{\mathrm{x}}(m)$ descriptors per image.

The problem is to decide if either a query $\mathbf{y} = \{\mathbf{y}^1, \cdots, \mathbf{y}^{J_{\mathrm{y}}}\}$ is related to some elements of the database or not. In the general case, $J_{\mathrm{y}} \neq J_{\mathrm{x}}(m)$. The system should produce a list of indices $\mathcal{L}(\mathbf{y})$ whilst ensuring that the correct index m is always on this list and an empty set, if the probe $\mathbf{y}$ is not related to any item in the database. The cardinality of this list or the number of retrieved indices on the the list is denoted as $|\mathcal{L}(\mathbf{y})|$.

System performance is evaluated by the probability of missing a correct item m on the list $\mathcal{L}(\mathbf{y})$ and the probability of falsely accepting an unrelated item m' as supposedly related to some item m in the database. This leads to the average list of accepted items $\mathbb{E}\{|\mathcal{L}(\mathbf{y})|\}$. Other parameters include memory storage, search complexity, security and privacy.[4, 18] In this paper, we will focus on the performance of the identification system for a given database size M, parameters of descriptors, their numbers $J_{\mathrm{x}}(m)$ and J_{y}, as well as targeting efficient search complexity based on inverted files.

The core idea behind the BOW systems consists in a representation of each image, described by a set of descriptors, by a fixed dimensional feature vector that is invariant to geometrical de-synchronization. The fixed-dimensional vector proves a robust and accurate approximation of image descriptors in terms of basis vectors or so-called *visual words*. In addition such a representation should ensure fast ϵ-NN or k-NN search. It is also closely related to the approximations or compression of the descriptors in the space of visual words that can be achieved by the *VQ (hard)-assignment*,[16] *source coding with refinement*[9] or *soft-assignment*.[11, 19]

The design of an equivalent codebook in terms of the best approximation of the descriptors is vital for the trade-off between memory storage, search complexity and accuracy. That is why the descriptors are stored in a compressed or approximated form $\mathbf{x}^i \rightarrow \hat{\mathbf{x}}^i$ with special indexing using hierarchical structures. However, to reveal the theoretical limits of the identification systems based on BOW, we will assume that the descriptors are uncompressed corresponding to the near perfect approximation that many state-of-the-art techniques strive for.[20]

For these reasons, we will consider the equivalent model shown in Figure 2 consisting of enrollment and identification via the equivalent codebook or alphabet $\mathcal{C}_{\mathrm{x}} = (\mathbf{x}^1, \cdots, \mathbf{x}^J)^T \in \mathbb{R}^{J \times L}$, where J is the number of codewords in the visual codebook. This equivalent codebook contains all unique descriptors, the composition of which gives a particular image $\mathbf{x}(m)$ representation. It should be pointed out that the representation of each image in terms of the equivalent codebook with an appropriate indexing structure makes it possible to obtain an efficient search.[9]

In this paper, we assume that each image $\mathbf{x}(m)$, $1 \leq m \leq M$, is represented by $J_{\mathrm{x}}(m)$ descriptors $\mathbf{x}^i(m)$, $1 \leq i \leq J_{\mathrm{x}}(m)$ as shown in Figure 2. The visual codebook $\mathcal{C}_{\mathrm{x}}$ contains all these descriptors labeled as $\mathbf{x}^j$, $1 \leq j \leq J$. In this case, the cardinality of the visual codebook is: $J = |\mathcal{C}_{\mathrm{x}}| = \sum_{m=1}^{M} J_{\mathrm{x}}(m)$. It should be remarked that in this part we assume that the visual codebook does not contain any clustering enabling efficient search nor any compression of descriptors. We are thus primarily interested in revealing the theoretical performance limit.

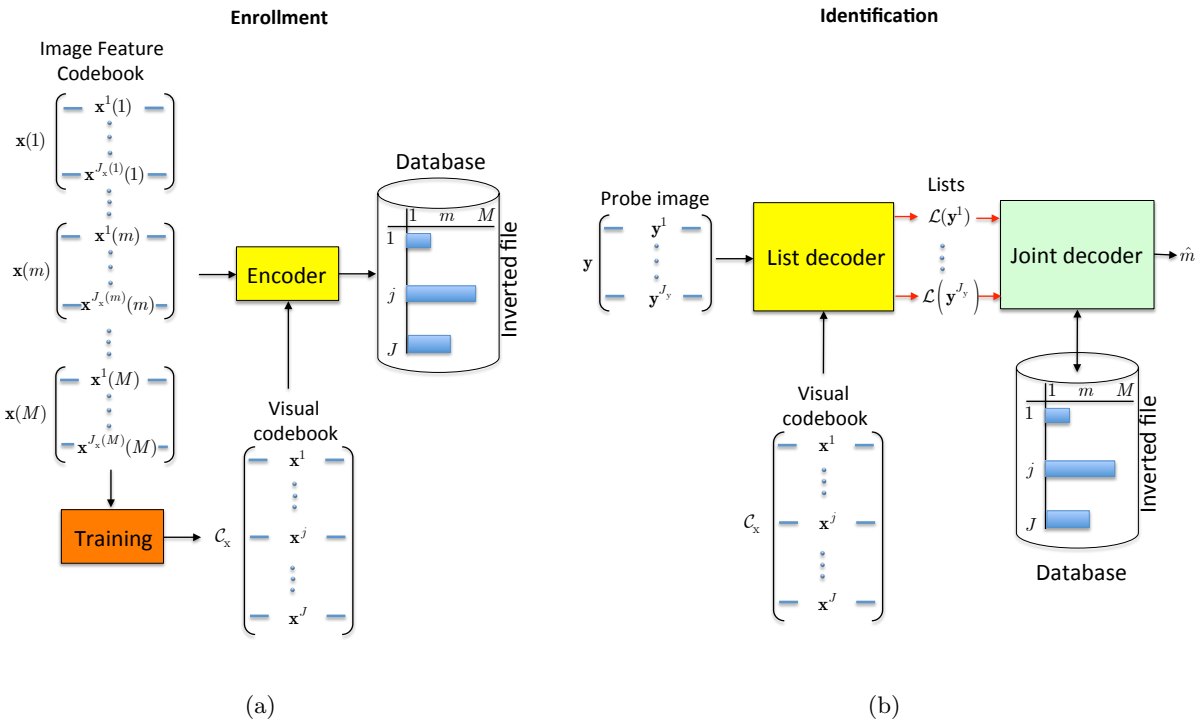


Figure 2. Enrolment (a) and identification (b) via the visual codebook.

At the end of the enrollment stage (Figure 2(a)), the encoder produces a database that is organised in the form of an inverted index file, i.e., each codeword with the index $j \in \{1, \cdots, J\}$ contains a list of images $m \in \{1, \cdots, M\}$ containing this particular visual word. At the identification stage, shown in (Figure 2(b)), the probe image $\mathbf{y}$, represented by its J_{y} descriptors, is presented to the identification system. The list decoder seeks all ϵ-NN or k-NN codewords $\mathbf{x}^j$, $j \in \{1, \cdots, J\}$ in the codebook $\mathcal{C}_{\mathrm{x}}$ thus producing J_{y} lists $\mathcal{L}(\mathbf{y}^k)$ for all

probe descriptors. It is assumed the the correct image index will be on the most lists with the high probability. The joint decoder observes these lists and test the database for the corresponding image indices containing the identified features and makes the final decision in favor of index $\hat{m}$ possessing the largest number of matched codewords.

3. STATISTICAL MODEL OF BOW CONTENT IDENTIFICATION

The statistical model of BOW content identification includes the definition of: (a) statistics of descriptors $\mathbf{x}^i$, (b) statistical observation model $p(\mathbf{y}^k|\mathbf{x}^i)$, (c) model of encoding/assignment, (d) model of decision making about the descriptor presence/absence and geometric consistency verification [†], and finally (e) model of global decision making or decoding.

Database of descriptors In this paper, we will assume that the local descriptors $\mathbf{x}^i \in \mathcal{X}^L$ are i.i.d. such as ORB following some distribution $\mathbf{X}^i \sim p(\mathbf{x}^i) = \prod_{n=1}^{L} p(x_n^i)$ [‡].

Statistical observation model The statistical observation model for the entire image is expressed in terms of the statistical model for the local descriptors:

$$p(\mathbf{y}|\mathbf{x}(m)) \equiv \prod_{k=1}^{J_\mathrm{y}} \prod_{i=1}^{J_\mathrm{x}(m)} p(\mathbf{y}^k|\mathbf{x}^i(m)), \tag{1}$$

which reduces to $p(\mathbf{y}|\mathbf{x}(m)) \equiv \prod_{k=1}^{J'} p(\mathbf{y}^k|\mathbf{x}^k(m))$ in the synchronised case, i.e., when the exact correspondence between the descriptors is known with $J' = \min\{J_\mathrm{x}(m), J_\mathrm{y}\}$.

The above probabilistic model can be also mapped into some metric space via $p(\mathbf{y}^k|\mathbf{x}^i(m)) \propto e^{-d(\mathbf{y}^k, \mathbf{x}^i(m))}$ assuming an exponential family of distortions, where $d(\mathbf{y}^k, \mathbf{x}^i(m))$ represents the distance between two descriptors.

Model of encoding/assignment In this paper, we will consider hard assignment to investigate the system performance under the minimum requested memory storage requirements[5, 9] [§]. The encoding matrix can be generally constructed as $\mathbf{C}_\mathrm{x}(m) = (\mathbf{c}_\mathrm{x}^1(m), \cdots, \mathbf{c}_\mathrm{x}^{J_\mathrm{x}(m)}(m)) \in \mathbb{R}^{J \times J_\mathrm{x}(m)}$, where each column $\mathbf{c}_\mathrm{x}^i(m)$ stands for the code representing the encoding of the descriptor $\mathbf{x}^i(m)$, $1 \le i \le J_\mathrm{x}(m)$ with respect to the visual codebook $\mathcal{C}_\mathrm{x}$. In the case of hard assignment, $\mathbf{C}_\mathrm{x}(m) \in \{0,1\}^{J \times J_\mathrm{x}(m)}$ with the elements $c_{\mathrm{x}_j}^i(m) = 1$ for $j : \mathbf{x}^j = \mathbf{x}^i(m)$ or zero-distance, i.e., $j \in \mathcal{L}(\mathbf{x}^i(m))$ with the list:

$$\mathcal{L}(\mathbf{x}^i(m)) = \{j \in \{1, \cdots, J\} : d(\mathbf{x}^j, \mathbf{x}^i(m)) = 0\}. \tag{2}$$

The encoding process based on the pooling is shown in Figure 3.

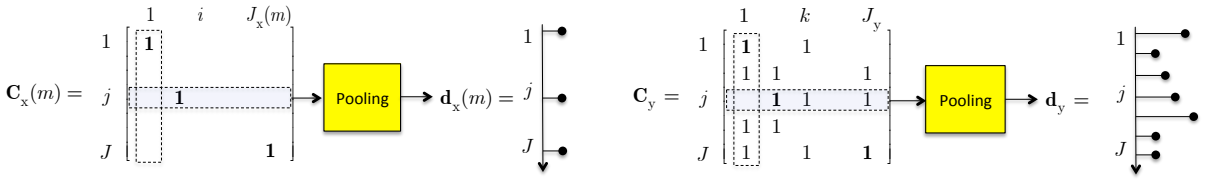


Figure 3. The block diagram of encoding/pooling at the enrollment and identification to produce a fixed-length representation of length J.

Given the case that the descriptors are matched without geometrical consistency, i.e., they are desynchronised, and there are generally a different number of descriptors in the enrolled image $J_\mathrm{x}(m)$ and probe J_y, *pooling* is

[†] We will investigate the upper theoretic limit assuming a perfect synchronisation between the descriptors.

[‡] One can consider different descriptors: local or global, sparse or dense, multilevel and binary. The i.i.d. assumption is not valid for SIFT descriptors which manifest a high correlation between elements.

[§] The hard/soft assignments represent a trade-off between the memory storage and decoding complexity.

used. To address this, there are two common types of pooling: *average*- and *max*- pooling. In the case of hard assignment at the enrollment stage they are equivalent. The enrolled fixed-length sparse code for the image m is $\mathbf{d}_{\mathrm{x}}(m) = (d^1_{\mathrm{x}}(m), \cdots, d^J_{\mathrm{x}}(m))^T \in \{0,1\}^J$ which is obtained as:

$$d^{av_j}_{\mathrm{x}}(m) = \sum_{i=1}^{J_{\mathrm{x}}(m)} c^i_{\mathrm{x}_j}(m), \tag{3}$$

in average-pooling and:

$$d^{max_j}_{\mathrm{x}}(m) = \max_{1 \le i \le J_{\mathrm{x}}(m)} c^i_{\mathrm{x}_j}(m), \tag{4}$$

in max pooling.

Model of decision making about the descriptor presence/absence Given a probe $\mathbf{y}$ consisting of J_{y} descriptors, the encoding matrix for the probe is defined as $\mathbf{C}_{\mathrm{y}} = (\mathbf{c}^1_{\mathrm{y}}, \cdots, \mathbf{c}^{J_{\mathrm{y}}}_{\mathrm{y}}) \in \{0,1\}^{J \times J_{\mathrm{y}}}$, with $c^k_{\mathrm{y}j} = 1$ for $j \in \mathcal{L}(\mathbf{y}^k)$ with the list $\mathcal{L}(\mathbf{y}^k) = \{j \in \{1, \cdots, J\} : d(\mathbf{x}^j, \mathbf{y}^k) \le \epsilon L\}$. This decoder corresponds to a bounded distance decoder (BDD) or ϵ-NN decoder which seeks all $\{\mathbf{x}^j\}$ NNs in the radius ϵL from the query descriptor $\mathbf{y}^k$, where ϵ is the threshold.

The performance of the descriptors is measured in terms of their ROCs defined by the probabilities of miss $P^D_M = \Pr\{d(\mathbf{x}^k, \mathbf{Y}^k) \ge \epsilon L\}$ and probability of false acceptance $P^D_F = \Pr\{d(\mathbf{x}^i, \mathbf{Y}^k) < \epsilon L\}$.

The fixed-length vector in the case of average pooling is defined as:

$$d^{av_j}_{\mathrm{y}} = \sum_{k=1}^{J_{\mathrm{y}}} c^k_{\mathrm{y}_j}, \tag{5}$$

and in the case of max pooling as:

$$d^{max_j}_{\mathrm{y}} = \max_{1 \le k \le J_{\mathrm{y}}} c^k_{\mathrm{y}_j}. \tag{6}$$

In following, we will compare max pooling and average pooling due to a common belief about the superior performance of max pooling.[11]

The statistics of matrix $\mathbf{C}_{\mathrm{y}}$ are completely defined by the probabilities of descriptor miss P^D_M and false acceptance P^D_F as defined above.

Model of global decision making or decoding The final decision is based on the list decoder that produces a list of possible candidates:

$$\mathcal{L}(\mathbf{y}) = \{m \in \{1, \cdots, M\} : t(m) \ge \tau J_e\}, \tag{7}$$

where:

$$t(m) = \mathbf{d}^T_{\mathrm{x}}(m)\mathbf{d}_{\mathrm{y}}, \tag{8}$$

stands for the similarity score between two vectors, for example the cosine distance, which is often used in the BOW systems, if the vectors are normalised by their norms $||\mathbf{d}_{\mathrm{x}}(m)||$ and $||\mathbf{d}_{\mathrm{y}}||$. We also define the equivalent length as $J_e = \min\{J_{\mathrm{x}}, J_{\mathrm{y}}\}$.

Remark: In the case when the correspondence between the descriptors from two images is established, one can estimate the upper bound on the system performance by evaluating the similarity between two matrices as $t(m) = \mathbf{C}_{\mathrm{x}}(m) \odot \mathbf{C}_{\mathrm{y}}$, where $\odot$ denotes the Frobenius inner product [¶].

[¶] In the synchronized case, the matrices are of the same size.

4. PERFORMANCE ANALYSIS: UNCOMPRESSED FEATURES

The overall system performance is evaluated by the probability of miss P_M, i.e., the correct m does not appear on the decoder's list under the hypothesis $\mathcal{H}_m$, $P_M = \Pr\{T(m) \leq \tau J_e | \mathcal{H}_m\}$ and by the probability of false acceptance P_F, i.e., an incorrect m' appears on the decoder's list under the hypothesis $\mathcal{H}_{m'}$, $P_F = \Pr\{T(m) > \tau J_e | \mathcal{H}_{m'}\}$, where τ is the threshold and J_e stands for the equivalent length. The average list size can be estimated as $\mathbb{E}\{|\mathcal{L}(\mathbf{y})|\} = MP_F$. In the case of unique identification, the list size is 1.

Without loss of generality, we will assume that the same number of descriptors is enrolled for all images, i.e., $J_{\mathrm{x}}(m) = J_{\mathrm{x}}$, which is a reasonable assumption for most identification systems where the enrollment is under the control.

The sufficient statistic in the case of max pooling and perfect synchronisation are:

$$T(m) \sim \begin{cases} \mathcal{B}(J_e, \theta(m)), & \text{for } \mathcal{H}_m, \\ \mathcal{B}(J_e, \theta(m')), & \text{for } \mathcal{H}_{m'}, \end{cases} \tag{9}$$

where for the max pooling: $\theta(m) = 1 - (1 - P_D^D)(1 - P_F^D)^{J_{\mathrm{y}}-1}$ and $\theta(m') = 1 - (1 - P_F^D)^{J_{\mathrm{y}}}$ and for the perfectly synchronised case: $\theta(m) = P_D^D$ and $\theta(m') = P_F^D$. The proof is given in Appendix A.

The performance of the content identification system is estimated based on the list decoder, which is characterised by the probability of miss:

$$\begin{aligned} P_M &= \Pr\{T(m) \leq \tau J_e | \mathcal{H}_m\} \\ &= \sum_{d=0}^{\tau J_e} \binom{J_e}{d} \theta^d(m)(1-\theta(m))^{J_e - d} \\ &\leq 2^{-J_e \mathcal{D}(\tau \| \theta(m))}, \end{aligned} \tag{10}$$

where $\mathcal{D}(\tau \| \theta(m))$ denotes the divergence and the probability of false acceptance is:

$$\begin{aligned} P_F &= \Pr\{T(m) > \tau J_e | \mathcal{H}_{m'}\} \\ &= \sum_{d=\tau J_e}^{J_e} \binom{J_e}{d} \theta^d(m')(1-\theta(m'))^{J_e - d} \\ &\leq 2^{-J_e \mathcal{D}(\tau \| \theta(m'))}, \end{aligned} \tag{11}$$

which results into the average list of candidates $\mathbb{E}\{|\mathcal{L}(\mathbf{Y})|\} = MP_F$. The threshold should satisfy $0 \leq \theta(m') < \tau < \theta(m) \leq 1$.

Using the notion of the identification rate as $R = 1/J_e \log_2 M$ defined for large J_e, one can target the condition $R \leq \mathcal{D}(\tau \| \theta(m'))$ to keep the list of retrieved candidates small.

The distribution of parameter $T(m)$ for the case of average-pooling is considered in Appendix B. To present the results in a tractable form, we use the Gaussian approximation that results in:

$$T(m) \sim \begin{cases} \mathcal{N}(\mu(m), \sigma^2(m)), & \text{for } \mathcal{H}_m, \\ \mathcal{N}(\mu(m'), \sigma^2(m')), & \text{for } \mathcal{H}_{m'}, \end{cases} \tag{12}$$

where $\mu(m) = J_e(P_D^D + (J_{\mathrm{y}} - 1)P_F^D)$ and $\mu(m') = J_e J_{\mathrm{y}} P_F^D$, $\sigma^2(m) = J_e(P_D^D(1 - P_D^D) + (J_{\mathrm{y}} - 1)P_F^D(1 - P_F^D))$ and $\sigma^2(m') = J_e J_{\mathrm{y}} P_F^D(1 - P_F^D)$ with $J_e = \min\{J_{\mathrm{x}}, J_{\mathrm{y}}\}$ and a new threshold $J_t = J_e J_{\mathrm{y}}$.

The corresponding performance under the average-pooling is [‖]:

[‖] We use a Gaussian approximation of Hamming distances assuming that the distances contribute only to non-negative values.

$$
\begin{aligned}
P_M &= \Pr\{T(m) \leq \tau J_t | \mathcal{H}_m\} \\
&\approx \int_0^{\tau J_t} \frac{1}{\sqrt{2\pi\sigma^2(m)}} e^{-\frac{(t-\mu(m))^2}{2\sigma^2(m)}} dt \\
&\approx Q\left(\frac{\mu(m) - \tau J_t}{\sigma(m)}\right), \qquad (13)
\end{aligned}
$$

$$
\begin{aligned}
P_F &= \Pr\{T(m) > \tau J_t | \mathcal{H}_{m'}\} \\
&\approx \int_{\tau J_t}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2(m')}} e^{-\frac{(t-\mu(m'))^2}{2\sigma^2(m')}} dt \\
&= Q\left(\frac{\tau J_t - \mu(m')}{\sigma(m')}\right), \qquad (14)
\end{aligned}
$$

with the the average list of candidates $\mathbb{E}\{|\mathcal{L}(\mathbf{Y})|\} = MP_F$ **.

In some applications, it is interesting to keep both probabilities of errors small. In this case, one can follow the strategy to minimise the maximum probability of error under optimal τ and ϵ defined as $(\hat{\tau}, \hat{\epsilon}) = \arg min_{\tau,\epsilon} \max\{P_M(\tau,\epsilon), P_F(\tau,\epsilon)\}$. We will investigate this problem numerically.

For technical reasons we will first fix ϵ and estimate τ. In the case of max pooling and perfect synchronisation, the above maximisation is achieved when $P_M(\tau,\epsilon) = P_F(\tau,\epsilon)$. The equality of (10) and (11) leads to the equality $\mathcal{D}(\tau\|\theta(m)) = \mathcal{D}(\tau\|\theta(m'))$ that yields:

$$
\hat{\tau} = \frac{\log \frac{1-\theta(m')}{1-\theta(m)}}{\log \frac{\theta(m)(1-\theta(m'))}{\theta(m')(1-\theta(m))}}, \qquad (15)
$$

From the equality of Q-function arguments in the equations (13) and (14), one can obtain for the average-pooling:

$$
\hat{\tau} = \frac{1}{J_t} \frac{\sigma(m')\mu(m) + \sigma(m)\mu(m')}{\sigma(m') + \sigma(m)}. \qquad (16)
$$

5. PERFORMANCE ANALYSIS: COMPRESSED FEATURES

The visual codebook compression targets three main objectives: (a) efficient search due to indexing based on inverted files, (b) satisfaction of memory storage requirements and (c) visual codebook learning based on the available training data ††. The different BOW architectures can be generalised in the scope of the hierarchical organisation of the visual codebook when ϵ-NN descriptors are clustered and represented by a common centroid. Then the probe is first tested versus all centroids and the nearest centroids are found. The descriptors belonging to the found clusters represented by these centroids are explored to find the ϵ-NN or k-NN descriptors.

Impact of compression on descriptors' ROC and overall BOW performance Disregarding a particular assignment technique, i.e., either hard- or soft-assignment, or source coding with refinement, the compression introduces the average descriptor distortion $D_c = \mathbb{E}\{d(\mathbf{X}^j, \hat{\mathbf{X}}^j)\}$, where $\hat{\mathbf{X}}^j$ stands for the compressed version of jth descriptor $\mathbf{X}^j$ and $d(.,.)$ denotes the Euclidian or Hamming distance.

The statistical model of compression is considered as a mapping $p(\hat{\mathbf{x}}^j|\mathbf{x}^j)$. In a high-rate compression regime, i.e., regime of small distortions, the additive model can be used $\mathbf{x}^j = \hat{\mathbf{x}}^j + \mathbf{z}^j$, where $\mathbf{z}^j$ is the compression noise, independent of $\hat{\mathbf{x}}^j$.[3] In the case of binary descriptors like ORB, the distortion D can also be interpreted as $\Pr\{X_i^j \neq \hat{X}_i^j\} \leq D$, where $x_i^j \in \{0,1\}$. If the binary descriptor is considered as Bernoulli's source with $\Pr\{X_i^j = 1\} = p$, then from rate-distortion theory it is known that $D(R_D) = H_2^{-1}(H_2(p) - R_D(D))$, for $0 \leq D \leq \min\{p, 1-p\}$, where R_D is the rate of the descriptor compression, and $H_2(.)$ denotes the binary

**The probability P_F can be found exactly as shown in Appendix B since $T(m) \sim \mathcal{B}(J_e J_y, P_D^D)$ under $\mathcal{H}_{m'}$.

††The last concerns mostly the CBIR systems. In content identification applications, the codebook design can be optimised for all available enrolled samples.

entropy.[3] For example, if the 256-bit-ORB descriptor with $p = 0.5$ is compressed to 64 bits, i.e., with the rate $R_D = 0.25$ bit/sample, the distortion is $D(R_D) = 0.2145$. Practically, the compression of the descriptor leads to extra degradation in the chain $\mathbf{Y}^j - \mathbf{X}^j - \hat{\mathbf{X}}^j$, when the compressed descriptors are stored instead of original ones. It results in the degradation of the ROC for the descriptors. Mostly it concerns the degradation for the geometrically consistent descriptors with $\Pr\{Y_i^j \neq \hat{X}_i^j\} = D * P_b$, where P_b is the probability of bit error, and for the geometrically inconsistent descriptors $\Pr\{Y_i^k \neq \hat{X}_i^j\} = D * 0.5 = 0.5$, where $D * P_b = D(1 - P_b) + P_b(1 - D)$. It means that the probability mass function for the geometrically consistent descriptors is closer to those of geometrically inconsistent ones and it has higher variance. Altogether, it leads to an increase of P_D^M and P_D^F.

The same arguments are valid for SIFT descriptors. We will consider a particular method of their compression based on the product-of-vector quantizer[9] and investigate the impact of compression on the ROC curves in Section 7.

Impact on indexing structure The indexing structure of modern BOW systems is mostly based on hierarchical clustering. The generalisation of different indexing structures addressing memory-complexity trade-off with the rate R close to the identification capacity $I(X^j; Y^j)$ is given in.[5] This study covered the case of unique decoding, i.e., identification, where each object is represented by a single descriptor in the codebook. On the contrary, in the BOW systems the rate $R > I(X^j; Y^j)$ and thus the system is not able of producing a unique decision. Instead, as considered in the previous Section, the list of the most likely candidates is produced based on list decoding. Moreover, the descriptors are additionally compressed in the BOW systems in contrast to the identification setup considered in[5] which reduces the identification rate $I(\hat{X}^j; Y^j) \leq I(X^j; Y^j)$. Therefore, this analysis is not directly applicable.

Summarising different methods of clustering based search, we can categorise them depending on: (a) method of generation of centroids, i.e., $p(\mathbf{c}^j|\mathbf{x}^j)$, $p(\mathbf{c}^j|\mathbf{y}^j)$ or $p(\mathbf{c}^j|\mathbf{x}^j, \mathbf{y}^j)$, (b) decoding, i.e., unique or list decoding and (c) covering principle. According to,[5] all these methods achieve the identification capacity under different memory-complexity relationships but generally the centroid generation based on $p(\mathbf{c}^j|\mathbf{x}^j, \mathbf{y}^j)$ with list encoding shows the best memory-complexity trade-off. It should be also pointed out that the main memory burden comes from the storage of uncompressed descriptors. Therefore, the optimal compression of descriptors is crucial.

In this paper, we only consider hard assignment with list decoding (or multi-query extension) for its reasonable memory-complexity trade-off. Therefore, in our model the centroids are assumed to be generated as $p(\mathbf{c}^j|\mathbf{x}^j)$ with no covering and list decoding similarly to.[9] The descriptors can be stored in their original form or compressed. This is reflected by the corresponding P_M^D and P_F^D.

In conclusion, besides the obvious advantage of more efficient memory storage, the compression of descriptors has several important negative implications. First, the compression of descriptors leads to the increase of P_M^D and P_F^D thus degrading the overall performance in terms of P_M and P_F. Secondly, the compression of descriptors extends the search region for the ϵ-NN centroid thus leading to an increased search complexity. Finally, it reduces the entropy of the descriptors therefore increasing the potential for attacks.

6. SECURITY CONSIDERATION

In this section, we consider the security of BOW search systems in terms of attacks targeting to deteriorate the system performance using prior information about the system design. Previous works,[10] have mostly considered empirical observations behind the removal/false creation of key points and modification of descriptors in the support regions without revealing the impact mechanisms of these modifications. In this paper, we will demonstrate for the first time, how these modifications are reflected on the overall system performance for different pooling approaches addressing the lack of geometric consistency. Moreover, we will also consider attacks against the indexing structure.

6.1. General considerations

It is natural to consider the security of BOW systems as a *game* between the system designer (defender) and attacker formulated as:

$$\max_{\mathcal{A}: d((\mathbf{x},\mathbf{y})) \leq D_A} \min_{\epsilon,\tau} \max\{P_M(\epsilon, \tau), P_D(\epsilon, \tau)\}, \tag{17}$$

where $\mathcal{A}$ denotes a class of attacker strategies with the bounded distortion D_A. The attacker strategy depends on prior knowledge on the system, its design and parameters, for both an informed and a blind attacker. In the general case, the system design might be known to the attacker but not any application specific parameters. However, the attacker can learn them by testing the system feedback on the modified images, similarly to the *sensitivity attack* in watermarking.[2] Therefore, the worst case attack scenario includes an informed attacker. In this case, all parameters of the system are fixed and the attacker is trying to reach his goal, i.e., by modifying the image $\mathbf{x}(m)$ to achieve its miss by the system or by producing a fake $\mathbf{x}'$ to achieve its false acceptance by the BOW system.

6.2. Informed attacks against the key points and descriptors

The overall system performance is fully determined by the statistics of $T(m)$ (9) and (12) under max pooling and perfect synchronisation, and average pooling, respectively. We suppose that the threshold τ is fixed in advance and the attacker aims to make the distributions of $T(m)$ under $\mathcal{H}_m$ and $\mathcal{H}_{m'}$ strongly overlapping. We will start with the uncompressed descriptors [‡‡].

Perfect synchronisation In this case, the modification of the number of key points/descriptors J_{y} in the probe will impact $J_e = \min\{J_{\mathrm{x}}, J_{\mathrm{y}}\}$ which decreases the distance between the distributions and leads to an increase of both the probability of miss (10) and false acceptance term in (14). The modification of descriptors for a fixed ϵ automatically alters $\theta(m) = P_D^D$ and $\theta(m') = P_F^D$, i.e., it also controls the means and variances of the binomial distributions. In particular, increasing P_F^D will move the distribution under the hypothesis $\mathcal{H}_{m'}$ closer to one under $\mathcal{H}_m$ thus producing a higher overall P_F.

Max pooling In the non-synchronised cases, the impact of attacks is more complex. The attacks against the key points are also directly reflected via $J_e = \min\{J_{\mathrm{x}}, J_{\mathrm{y}}\}$, i.e., the decrease of J_{y} reduces J_e. Contrarily to the perfectly synchronised case, J_{y} also has a significant impact on the error probabilities under both hypothesis $\theta(m) = 1-(1-P_D^D)(1-P_F^D)^{J_{\mathrm{y}}-1}$ and $\theta(m') = 1-(1-P_F^D)^{J_{\mathrm{y}}}$ especially in relationship with P_F^D. It is interesting to note that if the attacker removes all key points resulting in $J_{\mathrm{y}} = 1$, then $\theta(m) = P_D^D$ and $\theta(m') = P_F^D$ which coincides with the perfect synchronisation case. However, since $J_e = \min\{J_{\mathrm{x}}, 1\} = 1$, both hypothesis are hardly separable. On the contrary, the increase of J_{y} increases the impact of the false descriptors such that both $\theta(m)$ and $\theta(m')$ tend to 1 and the distributions are strongly overlapping. Therefore, the attacks against the key points are very important for max pooling. The attacks against the descriptors for the fixed J_{x} and J_{y} show the high sensitivity to the increase of P_F^D. If P_F^D increases, $\theta(m)$ and $\theta(m')$ converge to 1 which is similar to the impact due to the increase of J_{y}. The decrease of P_D^D to 0 also makes $\theta(m)$ and $\theta(m')$ indistinguishable.

Average pooling The average pooling corresponding to the non-synchronised case is also sensitive to attacks against the key points similar to max pooling. The removal of key points, i.e., decrease of J_{y}, reduces the distance between the means $\mu(m) = J_e(P_D^D + (J_{\mathrm{y}} - 1)P_F^D)$ and $\mu(m') = J_e J_{\mathrm{y}} P_F^D$. In the extreme case when $J_{\mathrm{y}} = 1$, $\mu(m) = P_D^D$ and $\mu(m') = P_F^D$ which coincides with the previous cases. The increase of J_{y} increases the non-distinguishability of the distributions which strongly manifests itself under the increased P_F^D . Similarly to max pooling, the decrease of P_D^D leads to the reduced distance between the means and the coincides of variances.

In the case of compressed descriptors, the statistics of $T(m)$ under $\mathcal{H}_m$ and $\mathcal{H}_{m'}$ are closer with respect to their uncompressed counterpart. Therefore, the attackers' job is even more simplified in terms of the required distortions.

6.3. Informed attacks against the indexing mechanism

The informed attacks against the indexing structure target the miss of a descriptor, i.e., the equivalent increase of P_M^D, leveraging information on the design of the indexing structure. Since most of the fast indexing methods are based on clustering, the informed attacker will target such a modification of a descriptor that it will not be covered by the radius of the ϵ-NN search. Equivalently, one can consider such modifications that the list of centroids of the predefined cardinality closest to the probe descriptor will not contain the corresponding enrolled descriptor. Contrarily to the blind attacker who introduces the blind distortions to all descriptors, the informed

[‡‡] We do not consider the attacks against key points based on which the descriptor is computed. The "removal" of key point obviously leads to the miss of descriptor with the probability 1.

attacker will induce a distortion specifically designed to move the descriptors outside the list decoding regions. Moreover, the attacker might proceed with the selective strategy using so-called *descriptor reliability*, i.e., the descriptors that are closer to the borders of Voronoi's regions are more vulnerable to flipping due to distortions or can be moved far way from the correct centroid under milder distortions.

7. RESULTS OF COMPUTER SIMULATION

Since the overall system performance is determined by the statistics of $T(m)$ which in turn is defined by P_M^D and P_F^D, we first investigated the typical ROC curves for SIFT and ORB descriptors shown in Figure 4 for the *copydays* database.[8] We have tested about 100'000 descriptors. As expected, SIFT produces better results, but is computationally heavy.

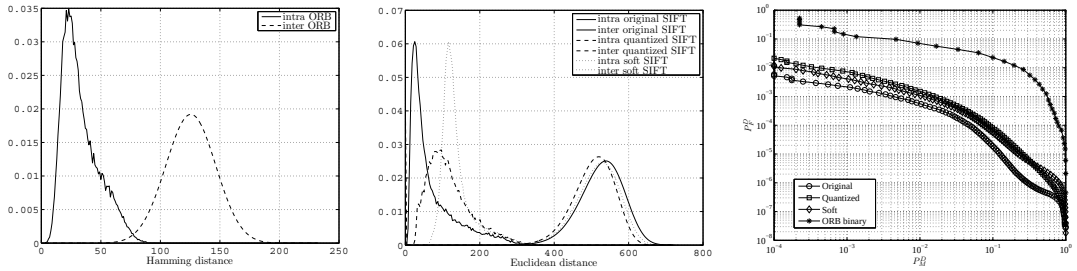
The experimental distributions of parameter $T(m)$ for the matched and non-matched pairs of descriptors are shown in Figures 4(a) and 4(b) for the ORB and SIFT descriptors, respectively. In addition, to investigate the impact of descriptor compression, we tested the product-of-vector quantizer proposed in[9] with a block size of 8 and 256 centroids in each block giving 64 bits in total per descriptor. The inter- and intra-class pdfs for the symmetric case, i.e., both the probe and enrolled descriptors are quantized, is referred to as *quantized SIFT*, and the asymmetric case, i.e., the probe descriptor is soft while the enrolled descriptor is quantized, is referred to as *soft SIFT*, and is shown in Figure 4(b) in comparison with the original SIFT pdfs. The ROCs for the ORB and SIFT descriptors are shown in Figure 4(c). The results are obtained for 100'000 matched pairs of descriptors. It should be remarked that the original SIFT descriptors demonstrate significantly better performance in comparison to ORB; at least in 2 orders of magnitude in terms of the P_F^D. However, ORB was about from 5 to 8 times faster in our experiments. The product-of-vector quantizer has demonstrated a remarkable performance. In particular, very good results were obtained for the quantized SIFT with 64 bits which outperforms 256 bit ORB. However, as it was mentioned, the computational burden of SIFT is an essential bottleneck in practical on-line applications. Therefore, there is a need for binary, fast and performant features which generate descriptors directly without any compression.

To validate the accuracy of the developed mathematical model, we experimentally tested the distribution of $T(m)$ in (9) and (12), for the average pooling, max pooling and synchronised case using the P_M^D and P_F^D for the ORB descriptors. It should be pointed out that it follows expectation that SIFT provides better results. However, our objective was to validate and exemplify the performance of practical systems that can be used in on-line applications such as on mobile phones. Therefore, we focused on the ORB descriptors leaving the comparison of different descriptors out of scope of this paper. That is why our objective was to, given the ROC curve of any descriptor, predict the performance of the BOW identification system.

Accordingly, the main hypotheses we wanted to test here were: (i) to confirm that max pooling is superior to average pooling as is commonly believed and (ii) to investigate the gap between pooling strategies that operate under the geometrical ambiguity with those of the perfectly synchronized case. For purely demonstrative purposes, we have chosen the operational point $P_M^D = 0.2954$ and P_F^D =0.0101 on the ORB ROC curve in Figure 4(c) and investigate the sufficient statistic $T(m)$ under $\mathcal{H}_m$ and $\mathcal{H}_{m'}$ as shown in Figure 5. The theoretical counterparts follow the experimental results remarkably well. The results were achieved in 10 million experiments for each type of pooling. The number of descriptors in the enrolled images and probe was identical and equals 500.

To obtain the objective performance of BOW systems, we investigated the ROC curves for the above operational point. In addition, it is interesting to highlight the impact of the different relationship between J_{x} and J_{y} on the overall system performance, which is why we considered four cases: (a) $J_{\mathrm{x}} = J_{\mathrm{y}} = 50$ to simulate some practical situations, (b) $J_{\mathrm{x}} = J_{\mathrm{y}} = 500$ to investigate the impact of an increased number of descriptors, (c) $J_{\mathrm{x}} = 500$ and $J_{\mathrm{y}} = 50$ to study the impact of cropping in the probe and (d) $J_{\mathrm{x}} = 50$ and $J_{\mathrm{y}} = 500$ to highlight the impact of collage or probe acquisition with some background interfering objects/features. The results of this study are shown in Figure 6 and lead to the following conclusions. First, the overall performance of max pooling and average pooling for this operational point is comparative (Figure 6(a)). Second, as expected the synchronised case represents the lower bound in all tests and there is a considerable gap with average and max pooling. This signifies the importance of geometrical synchronisation. We strongly believe that the BOW methods should strongly benefit from local geometrical consistency verification at the early stages of descriptor matching in contrast to the existing architecture presented in Figure 1, where the geometrical verification is performed in the last

stage. Third, the increase of the number of descriptors from $J_x = J_y = 50$ to $J_x = J_y = 500$ drastically increases the performances of the synchronised case (Figure 6(b)). This is in accordance with detection and information theory. The increase of the number of descriptors in the synchronised case is equivalent to the creation of one super descriptor, or vector, with a total length of 500x256 bits in the ORB case. Contrary, the performance of average pooling and max pooling drops. This is due to the fact that the BOW system generates a lot of false matches for the non-synchronised descriptors and the increase of their number leads to the masking of and interference with, the correct matches. In addition, the drop in performance for max pooling is higher. This is due to the fact that average pooling generates a sort of soft information about the number of matches in the considered model while max pooling represents a binary decision rule that just declares the presence or absence of at least one match. Forth, the asymmetric case of $J_x = 500$ and $J_y = 50$, that represents a sort of cropping, does not show any difference with the case of $J_x = J_y = 50$ (Figure 6(c)). Fifth, the asymmetric case of $J_x = 50$ and $J_y = 500$ represents a case with many false descriptors in the probe image (Figure 6(d)). The synchronised case shows good immunity to that. However, average pooling and max pooling manifest a drastic drop in their performance. This is due to the overwhelming amount of false matching. That is why the sole matching of descriptors is not sufficient without additional information on their geometric consistency.



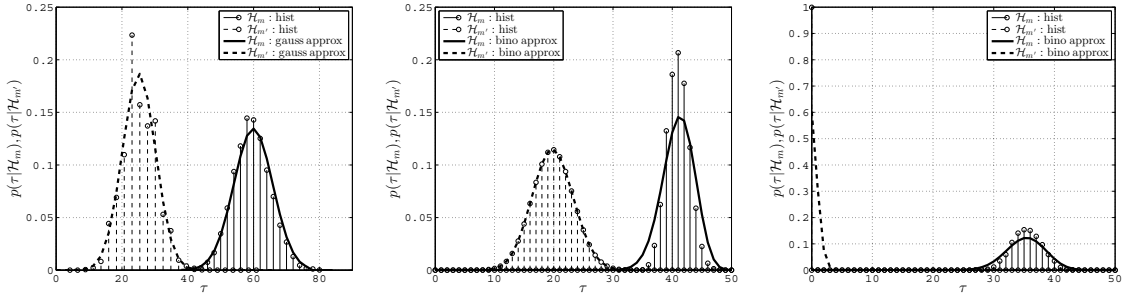
(a) (b) (c)

Figure 4. Typical performance of descriptors for matched and non-matched descriptors under scaling 0.5, rotation 10^0 and JPEG 75: (a) histograms for ORB (uncompressed-uncompressed) descriptors, (b) histograms for SIFT (uncompressed-uncompressed, uncompressed-compressed, compressed-compressed) and (c) ROCs for ORB and SIFT descriptors.

Finally, the overall performance of BOW content identification system is summarised in terms of the min $\max\{P_M(\tau,\epsilon), P_F(\tau,\epsilon)\}$ as a function of ϵ and τ in Figure 7. It is interesting to note that the system has a global minimum for the optimal pair of the thresholds τ, ϵ. We investigated the optimal pair of τ, ϵ leading to the minimisation of maxim error (Figure 7(a)). It should be noted that the experimental thresholds are very accurately predicted by the derived formulas (15) and (16). Therefore, the optimisation of BOW systems is straightforward. It is also remarkable that in our model max pooling did not demonstrate any advantage over average pooling, contrarily to common believe. Similar performance to max pooling was observed under the proper ϵ and τ. Interestingly, both methods have the same global minimum for the same ϵ=0.25 (Figure 7(b)). However, the optimal τ valuess are different for max pooling and average pooling (Figure 7(a)). Therefore, if one specifies the descriptor, our model suggests a set of optimal parameters under average and max pooling to optimize overall system performance.

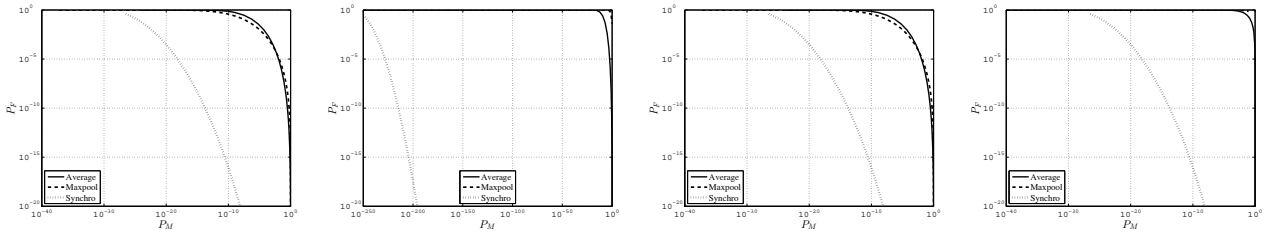
8. CONCLUSION

In this paper, we introduced a simple and tractable model of BOW content identification systems. The model is based on hard assignment and unique non-repeatable descriptors for each image. Average pooling is compared to max pooling where we have demonstrated the equivalence of both methods under the optimised system parameters. The importance of geometrical information for descriptor matching is demonstrated. We plan to (i) extend the proposed model to repetitive descriptors, (ii) find the maximum number of items M leading to a non-exponential list size of candidates and and finally (iii) investigate the impact of soft assignment on overall system performance. In this paper, we only highlighted the security issues of BOWs systems with an informed attacker. The proposed model can also be the basis for a theoretical investigation of different attacks against the BOW systems.



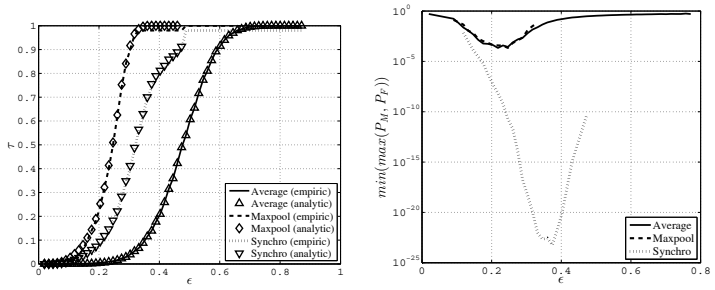
(a) (b) (c)

Figure 5. Distributions of similarity score for the ORB descriptor computed experimentally with the corresponding theoretical predictions: (a) average pooling, (b) max pooling and (c) synchronised case.



(a) (b) (c) (d)

Figure 6. Impact of the number of descriptors J_x and J_y on the ROC for the ORD descriptor for the descriptor operational point determined by $P_M^D = 0.2954$ and $P_F^D = 0.0101$ for average-, max pooling and synchronised case: (a) $J_\mathrm{x} = J_\mathrm{y} = 50$, (b) $J_\mathrm{x} = J_\mathrm{y} = 500$, (c) $J_\mathrm{x} = 500$, $J_\mathrm{y} = 50$ and (d) $J_\mathrm{x} = 50$, $J_\mathrm{y} = 500$.



(a) (b)

Figure 7. The optimal set of thresholds ϵ and τ for the ORB descriptor computed according to $(\hat{\tau}, \hat{\epsilon}) = \arg min_{\tau,\epsilon} \max\{P_M(\tau, \epsilon), P_F(\tau, \epsilon)\}$ for the average pooling, max pooling and synchronised cases: (a) the relationship between the thresholds and (b) the achievable error in the function of ϵ for the optimal τ. Remarkably, the average pooling and max polling achieve the same error for the optimised parameters.

9. ACKNOWLEDGMENT

This paper was partially supported by SNF grant 1200020-146379 and the Swiss-Polish project.

10. APPENDIX A

In this appendix we derive the distribution of parameter $T(m)$ under the hypotheses $\mathcal{H}(m)$ and $\mathcal{H}(m')$ for the max pooling. The equation (8) can be rewritten as:

$$t(m) = \mathbf{d}_{\mathrm{x}}^T(m)\mathbf{d}_{\mathrm{y}} = \sum_{j=1}^{J_e} d_{\mathrm{x}_j}(m) d_{\mathrm{y}_j} = \sum_{j'=1}^{J_e} d_{\mathrm{y}_{j'}}, \tag{18}$$

where $J_e = \min\{J_{\mathrm{x}}, J_{\mathrm{y}}\}$. The parameter $T(m)$ represents a sum of J_e i.i.d. Bernoulli random variables $D_{\mathrm{y}_{j'}} = D_{\mathrm{y}}$ with the probability of success $\Pr\{D_{\mathrm{y}} = 1|\mathcal{H}(m')\}$ under the hypothesis $\mathcal{H}(m')$. Max pooling makes a positive decision on the presence of the descriptor, if at least one positive decision out of J_{y} is observed, which is defined as:

$$\theta(m') = \Pr\{D_{\mathrm{y}} = 1|\mathcal{H}(m')\} = \Pr\{\bigcup_{k=1}^{J_{\mathrm{y}}} E_F\} = 1 - \Pr\{\bigcap_{k=1}^{J_{\mathrm{y}}} \bar{E}_F\} = 1 - \prod_{k=1}^{J_{\mathrm{y}}}(1 - P_F^D) = 1 - (1 - P_F^D)^{J_{\mathrm{y}}}, \tag{19}$$

where E_F denotes an event consisting in a false acceptance and $\bar{E}_F$ denotes non-E_F, i.e., the correct rejection.

Similarly, under the hypothesis $\mathcal{H}(m)$, max pooling will make a positive decision, if there is at least one match coming from the true descriptor or from $J_{\mathrm{y}} - 1$ falsely matched descriptors:

$$\theta(m) = \Pr\{D_{\mathrm{y}} = 1|\mathcal{H}(m)\} = \Pr\{E_c \cup \bigcup_{k=1}^{J_{\mathrm{y}}-1} E_F\} = 1 - \Pr\{\bar{E}_c \cap \bigcap_{k=1}^{J_{\mathrm{y}}-1} \bar{E}_F\} = 1 - (1 - P_D^D)(1 - P_F^D)^{J_{\mathrm{y}}-1}, \tag{20}$$

where E_c and $\bar{E}_c$ denote the correct match with probability P_D^D and miss (complement to the correct match) with probability $(1 - P_D^D)$, respectively.

The summation of J_e independent Bernoulli random variables in (18) results in a Binomial random variable with the distributions $T(m) \sim \mathcal{B}(J_e, \theta(m'))$ under $\mathcal{H}_{m'}$ and $T(m) \sim \mathcal{B}(J_e, \theta(m))$ under $\mathcal{H}_m$.

In a similar way, one can obtain the distortion of parameter $T(m)$ for the synchronised case. The only difference consists in the fact that there is no summation over all J_{y}. This results in $\theta(m') = P_F^D$ and $\theta(m) = P_D^D$.

It is also interesting to remark that the term $\Pr\{E_c \cup \bigcup_{k=1}^{J_{\mathrm{y}}-1} E_F\}$ in (20) can be upper bounded as:

$$\theta(m) = \Pr\{E_c \cup \bigcup_{k=1}^{J_{\mathrm{y}}-1} E_F\} \leq \Pr\{E_c\} + \Pr\{\bigcup_{k=1}^{J_{\mathrm{y}}-1} E_F\} = P_D^D + (1 - \Pr\{\bigcap_{k=1}^{J_{\mathrm{y}}-1} \bar{E}_F\}) = P_D^D + (1 - (1 - P_F^D)^{J_{\mathrm{y}}-1}), \tag{21}$$

based on the union bound. Comparing (21) with $\theta(m) = P_D^D$ under the perfect synchronisation, it is easy to remark the impact of max pooling leading to the false matching term $(1 - (1 - P_F^D)^{J_{\mathrm{y}}-1})$. The higher P_F^D, the higher the contribution of this term to the overall probability of making a false decision on the descriptor presence.

11. APPENDIX B

In this part, we will consider the statistics of $T(m)$ under average pooling. The main difference with max pooling consists in the averaging of all positive matches in the variable $d_{\mathrm{y}_{j'}}$ in (18) rather than just indicating the positive success in any of J_{y} outcomes.

Under the hypothesis $\mathcal{H}_{m'}$, when only false matches can lead to the success event, there is a sum of J_{y} Bernoulli random variables with success probability P_F^D. Therefore, the summation of J_e independent Binomial random variables $D_{\mathrm{y}_{j'}} \sim \mathcal{B}(J_{\mathrm{y}}, P_F^D)$ in (18) results in a Binomial random variable with the distribution $T(m) \sim \mathcal{B}(J_e J_{\mathrm{y}}, P_D^D)$ under $\mathcal{H}_{m'}$.

Under the hypothesis $\mathcal{H}_m$, the random variable $d_{\mathrm{y}_{j'}}$ in (18) contains one correct match event with probability P_D^D and $J_{\mathrm{y}}-1$ matches with probability P_F^D. Unfortunately, it is not tractable to present the resulting distribution containing a sum of Binomial random variables with different dimensionalities and probabilities. Therefore, we will use the approximation of the sum of independent random variables by the Gaussian distribution according to the central limit theorem resulting in $T(m) \sim \mathcal{N}(\mu(m), \sigma^2(m))$, where $\mu(m) = J_e(P_D^D + (J_{\mathrm{y}} - 1)P_F^D)$ and $\sigma^2(m) = J_e(P_D^D(1 - P_D^D) + (J_{\mathrm{y}} - 1)P_F^D(1 - P_F^D))$.

Similarly, one can approximate the Binomial distribution under the distribution $\mathcal{H}_{m'}$ as $T(m) \sim \mathcal{N}(\mu(m'), \sigma^2(m'))$ where $\mu(m') = J_e J_{\mathrm{y}} P_F^D$ and $\sigma^2(m') = J_e J_{\mathrm{y}} P_F^D(1 - P_F^D)$.

REFERENCES

1. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *In ECCV*, pages 404–417, 2006.
2. P. Comesana, L. Perez-Freire, and F. Perez-Gonzalez. The blind newton sensitivity attack, 2006.
3. T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, 1991.
4. Farzad Farhadzadeh, Sviatoslav Voloshynovskiy, and Oleksiy J. Koval. Performance analysis of content-based identification using constrained list-based decoding. *IEEE Transactions on Information Forensics and Security*, 7(5):1652–1667, 2012.
5. Farzad Farhadzadeh, Frans M.J. Willems, and Sviatoslav Voloshynovskiy. Fundamental limits of identification: Identification rate, search and memory complexity trade–off. In *IEEE International Symposium on Information Theory (ISIT)*, Istanbul, Turkey, July 7–12 2013.
6. Teddy Furon, Hervé Jégou, Laurent Amsaleg, and Benjamin Mathon. Fast and secure similarity search in high dimensional space. In *IEEE International Workshop on Information Forensics and Security*, Guangzhou, China, 2013.
7. Bernd Girod, Vijay Chandrasekhar, David M Chen, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, Gabriel Takacs, Sam S Tsai, and Ramakrishna Vedantham. Mobile visual search. *Signal Processing Magazine, IEEE*, 28(4):61–76, 2011.
8. Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In Andrew Zisserman David Forsyth, Philip Torr, editor, *European Conference on Computer Vision*, volume I of *LNCS*, pages 304–317. Springer, oct 2008.
9. Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):117–128, 2011.
10. O. Koval, S. Voloshynovskiy, P. Bas, and F. Cayre. On security threats for robust perceptual hashing. In *Proceedings of SPIE Photonics West, Electronic Imaging / Media Forensics and Security XI*, San Jose, USA, 2009.
11. Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2486–2493. IEEE, 2011.
12. D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.
13. Benjamin Mathon, Teddy Furon, Laurent Amsaleg, and Julien Bringer. Secure and Efficient Approximate Nearest Neighbors Search. In ACM, editor, *1st ACM Workshop on Information Hiding and Multimedia Security*, IH & MMSec '13, pages 175–180, Montpellier, France, June 2013.
14. Pierre Moulin. Statistical modeling and analysis of content identification. In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–5. IEEE, 2010.
15. Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571. IEEE, 2011.
16. Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.
17. Avinash L Varna and Min Wu. Modeling and analysis of correlated binary fingerprints for content identification. *Information Forensics and Security, IEEE Transactions on*, 6(3):1146–1159, 2011.
18. S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, and T. Holotyak. Information-theoretical analysis of private content identification. In *IEEE Information Theory Workshop, ITW2010*, Dublin, Ireland, Aug.30-Sep.3 2010.
19. Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.
20. Christian Wengert, Matthijs Douze, and Hervé Jégou. Bag-of-colors for improved image search. In *ACM Multimedia*, Scottsdale, United States, October 2011.
21. Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.