Mobile visual object identification: from SIFT-BoF-RANSAC to SketchPrint

Sviatoslav Voloshynovskiy, Maurits Diephuis and Taras Holotyak

Computer Science Department, University of Geneva
7, Route de Drize, CH-1227 Carouge (GE), Switzerland

ABSTRACT

Mobile object identification based on its visual features find many applications in the interaction with physical objects and security. Discriminative and robust content representation plays a central role in object and content identification. Complex post-processing methods are used to compress descriptors and their geometrical information, aggregate them into more compact and discriminative representations and finally re-rank the results based on the similarity geometries of descriptors. Unfortunately, most of the existing descriptors are not very robust and discriminative once applied to the various contend such as real images, text or noise-like microstructures next to requiring at least 500-1'000 descriptors per image for reliable identification. At the same time, the geometric re-ranking procedures are still too complex to be applied to the numerous candidates obtained from the feature similarity based search only. This restricts that list of candidates to be less than 1'000 which obviously causes a higher probability of miss. In addition, the security and privacy of content representation has become a hot research topic in multimedia and security communities. In this paper, we introduce a new framework for non-local content representation based on *SketchPrint descriptors*. It extends the properties of local descriptors to a more informative and discriminative, yet geometrically invariant content representation. In particular it allows images to be compactly represented by 100 SketchPrint descriptors without being fully dependent on re-ranking methods. We consider several use cases, applying SketchPrint descriptors to natural images, text documents, packages and micro-structures and compare them with the traditional local descriptors.

Keywords: mobile visual search, bag-of-word, local descriptors, RANSAC, image features, mobile phones.

1. INTRODUCTION: THE APPLICATION

Visual identification of physical objects using a consumer mobile device has many applications in the security domain, for example in anti-counterfeiting (Figure 1), product tracking and tracing. It also represents a powerful tool for interaction with the physical world and creates new possibilities for augmented reality, e-commerce and market analysis. For example, the recognized object can be connected to the brand owner web site to provide more details about the product, indicate the network of stores where the product can be found with some promotion or sale or advise trusted partners where the authenticity of the product is ensured. In addition, the recognized products can be shared with friends in social networks or immediately verified for its authenticity with optional advices on how to proceed in the case of suspicious objects. In return brand owners and manufacturers might obtain very valuable information for market analysis and development of proper marketing strategies. These applications include but are not limited to reliable identification of diverse physical objects such as packages, watches, digital devices and chips, printed text documents, CDs, books, logos, object labels or etiquettes etc., such that each individual sample may be linked to their corresponding electronic identifier. In this respect, the visual identification of physical objects represents an interesting alternative to barcodes and more recently to digital watermarks[1] used in the retail industry. Both barcodes and digital watermarks assume that each object is assigned a unique index $m \in \{1, \cdots, M\}$ that is encoded into a form of barcode or watermark, generally referred as *marking*, and properly "embedded" into the object by printing or laser engraving. However, this might face some constraints and technical difficulties for certain applications where the bar codes or watermarks are not acceptable (watch industry), require modification of manufacturing processes, not always suitable for objects with a dominating white background (packages) or do not contain enough features for marking (text documents), and

Corresponding author: S. Voloshynovskiy, e-mail: svolos@unige.ch

Figure 1: An example of package identification based on a mobile phone. The identified object is immediately connected to the desired app providing for example information about the current sales or trusted vendours whose nearby geographical locations are indicated on the Google map (right picture).

are easily cloneable. The cloneability of this marking represents a major concern for the security industry. It is also likely that each application with its own content (images, text, logos, etc.) and manufacturing process might require its own marking technology which makes the marking non-universal and requires considerable customization engineering efforts.

At the same time, the identification of physical objects based on existing visual features does not require any modification of the objects and can be applied to various types of objects which make this process universal, non-intrusive and back-compatible to existing objects produced in past. It is also important to mention that the identification features can be acquired at different scales and include visible features that are observable by the naked eye or micro-features that are extracted at μm scale.[2] This is even possible with a non modified handheld mobile device.[3,4] The algorithm must be able to identify physical objects reliably even though the acquired images naturally suffer from geometrical projective transformations, non-linear distortions and varying illumination conditions.

Finally, it is important to remark that the considered mobile identification concerns the identification of a particular unique object rather than the semantic classification to a particular type or class.

2. RELATED WORK AND CHALLENGES

Our goal is to develop a framework for efficient visual object identification with a handheld device such as a mobile phone, even though these devices have limited optical resolution, computational power and bandwidth. These technical limitations create certain algorithmic constraints that should be carefully addressed in mobile applications. This concerns both possibilities when the feature extraction and identification are performed directly on mobile phones or only extracted features are sent to the server which perform the identification. The latter looks to be the most likely scenario for the large-scale, when the number of objects M is in the order of millions, and security applications, when the authentic object features are not disclosed and the identification engine is not public.

2.1 Existing identification architectures

Without pretending to be exhaustive in our overview of visual object recognition based on descriptors*, we can differentiate several existing architectures that can be used for object identification. The overall goal of identification consists in the estimation of the index $\hat{w}$, assigned to the object at enrollment stage, based on a probe image $\mathbf{y}$ †.

It is commonly known that local descriptors accompanied by spatial information on their originating geometrical positions represent a powerful tool for image recognition and alignment.[5] Disregarding the complexity, one can achieve excellent identification results using RANSAC based matching as shown in Figure 2a. The drawback of this approach is the need to perform brute force matching which could be unfeasible for real life applications.

*We do not address recognition systems based on convolutional neural networks.

†In cases when unique identification is not possible, a short list of 5-10 objects that have very similar features is produced and the end user is free to choose which object was presented for identification from this list

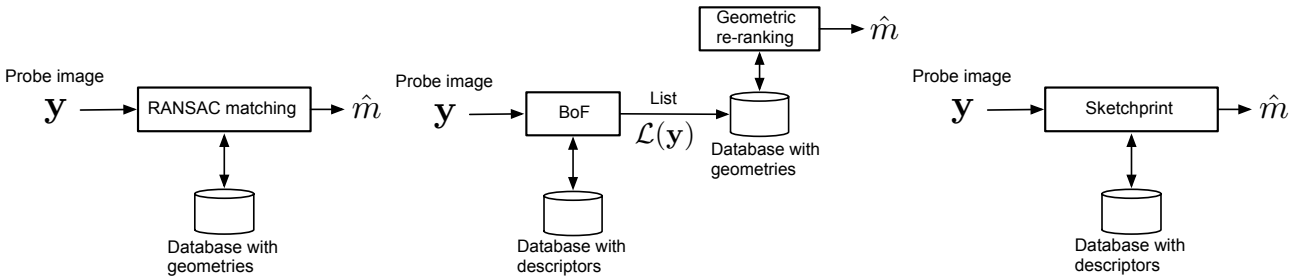

Figure 2: Identification architectures based on: (a) geometric matching using RANSAC, (b) BoF with geometric re-ranking and (c) SketchPrint descriptors without geometric re-ranking.

Partly to ease the computational burden, there exists a family of methods based on the bag-of-features (BoF) model applied to local image descriptors[6] or local geometric configurations of descriptors, also known as *geometric hashing*[7] (Figure 2b). The main idea behind this approach is to reduce the search space of size M for the RANSAC based geometric matching by producing a short list $\mathcal{L}(\mathbf{y})$ of images of a size not exceeding a 1'000 containing similar descriptors to the probe. The main advantage of BoF based encoding using a histogram of descriptors is it's relatively high accuracy for retrieval and acceptable complexity and memory storage for moderately sized databases. Further improvements were achieved via pyramid based BoF frameworks,[8] advanced aggregation strategies such as VLAD[9] and methods such as Fisher vectors.[10] The latter extends the BoF frameworks by integrating higher order statistics near attained BoF centroids. Another direction of research targets memory and computational complexity by compressing local descriptors to very short binary codes along with their geometrical coordinates within the originating images.[11,12] Identification accuracy, however, suffers from such optimisations.

Finally, our architecture, shown in Figure 2c, does not contain any geometric re-ranking stage and is based solely on robust and discriminative descriptors. This architecture represents our targeted scenario that will be considered in this paper. Since content descriptors form a basis of the considered approaches, we will consider the main shortcomings of existing descriptors.

2.2 Existing descriptors

Most local descriptors are only computed in the vicinity of special points known as key-points to cope with memory-complexity issues [‡]. The key-points are geometrically robust features that are often computed in salient image regions such as corners for example. Once the key-points are found the local descriptor is a function of its neighbouring pixels aligned with respect to some unique parameter of this key-point such as for example the major gradient. The descriptor function is often computed either as some relationship between the pixels in this neighbourhood (BREEF, BRISK, ORB)[12,13] or as function of local gradients or its histograms (SIFT, SURF)[14] or its VQ quantized counterpart (CHoG).[11] It should be pointed out that despite of the different complexity and the required number of bits, the SIFT descriptor remains one of the best in terms of its performance, amongst all of the above descriptors.

Without pretending to be exhaustive in our analysis, one can highlight the main shortcomings of local descriptors by considering the key-point detectors and descriptors separately.

The main problems of existing key-point detectors consist in:

- very weak robustness of existing methods to scaling and changing of lightening conditions which results in high probabilities of birth and death of key-points;
- very low precision of location key-points;

[‡]We do not consider dense descriptors here due to their huge memory-complexity

- absence of measures for the selection or ordering of reliable key-points for situations when the number of key-points should be restricted to some controllable quantity;
- very unstable performance on images that do not contain sharp corners such as random structures and uniform textures.

At the same time, most of the existing descriptors are characterized by:

- weak discriminative power since the descriptors are computed around corners, the entropy of these descriptors is relatively low; this especially concerns text documents where most of the characters will have very close descriptors repeated over the same document and random structure images;
- low robustness to geometric transforms that lead to high Euclidean/Hamming distances between corresponding descriptors.

2.3 Challenges and objectives

Taking the weaknesses into account of existing key-point detectors and low discriminative power of local descriptors, most existing identification systems try to benefit from the redundancy of these descriptors over the image, i.e., practically requiring about 500 descriptors for natural images and about 2000-3000 for random structures and text images. This unavoidably leads to an increase of memory usage and necessitates deploying advanced descriptor compression, aggregation and compression strategies, targeting about 64 bits per descriptor.

In addition, the geometrical positions or relationships between these local descriptors should be encoded and stored to ensure (approximate) geometric re-ranking at the final stage to prune the identification results. Such an approach represents the main stream approach in computer vision and content identification.

Therefore, in this paper, we present another concept that aims at resolving two main shortcomings of existing methods by: (a) using a non-local content description and (b) avoiding any geometric re-reranking as suggested by the architecture shown in Figure 2c.

In turn, we will demonstrate that a non-local content description is more informative in comparison to a local content one, robust to signal processing modifications and gives reasonable performance in face of typical geometric distortions as found in mobile imaging applications.

The fact that the content description is more discriminative next to implicitly encoding geometry, avoids the necessity of re-ranking at the final stage.

3. SKETCHPRINT: DEFINITION AND IMPLEMENTATION

In this section, we consider an identification system with the architecture presented in Figure 2c. At the base of the proposed framework is the *SketchPrint* descriptor[§]. We consider a sketch to be a read out of a signal between any reference system defined by two key-points. This signal should be properly processed to ensure its invariance to the signal processing distortions and geometric transformations.

The SketchPrint consists of four main stages:

- key-points detection;
- key-points filtering;
- SketchPrints extraction;
- SkechPrints filtering.

[§]The name SketchPrint comes from the *sketch*ing as a way of extracting features and content finger*printing*.

3.1 Key-points detection

In fact, the SketchPrint can be used by any known robust key-point detection method enabling stable and repeatable detection of the same key-points under the envisioned family of geometric transforms and image processing distortions. However, there are also three specific requirements to the key-point detector for SketchPrint that should be considered while choosing an appropriate method:

- *informativness* (**R1**): to ensure the most informative representation of the image, the SketchPrint descriptor should extract those cross-section signals in the regions that are the most informative (distinctive) for identification. Practically, it means that, for example, the flat regions or regions with small content variabilities are not favourable for SketchPrint. In contrast, the regions possessing high signal variability or high entropy are more preferable. This will clearly help distinguish one image from the other with the smallest number of descriptors.
- *robustness* (**R3**): to withstand various geometric distortions the key-points in the SketchPrint algorithm should satisfy several requirements for robustness:

  (**R3.1**) (repeatability of key-points): The first requirement concerns the high repeatability of key-points after the envisioned distortions. This requirement is very important for SketchPrint and high repeatability is even more crucial in comparison to traditional descriptors based on a single key-point. The reason is that SketchPrint is defined by a system of two key-points in contrast to local descriptors where just a single key-point is used to define the position for the descriptor. That is why the key-points for SketchPrint should be even more robust. If the probability of observing the same key-point in the original and the distorted images is P_D^p than due to the independence of the key-points, the probability of observing two key-points is a product $P_D^p P_D^p$. Practically, it means that the system of two key-points is less stable to distortions than just a single one. Unfortunately, up to our best knowledge the existing key-point detectors have a probability of observing the same point of around 0.85 or even significantly less which would make SketchPrint extremely vulnerable to the loss of key-points. Therefore, in the next section we will present a solution based on the redundancy of local key-points produced by certain key-point detection methods.

  (**R3.1**) (robustness to non-linear geometric distortions): The second requirement concerns the robustness to strong non-affine, nonlinear geometric and lens distortions which might occur in mobile imaging applications. To cope with this requirement, we will assume that these distortions, that are even difficult to be characterized mathematically, are approximated by a locally affine transform. This assumption was successfully validated and used in the digital watermarking to withstand a random banding attack. In the case of SketchPrint, it means that, if the key-points are chosen to be not far from each other and confirm this assumption, all concerned distortions will reduce to just scaling. In turn to be invariant to scaling, the cross-section between two key-points can always be re-scaled to a fixed length thus creating perfect invariance. We will address this requirement in the key-point filtering section.

It is worth mentioning again that our overall goal is to obtain a very compact, informative and robust content representation with less then 50-100 SketchPrint descriptors per image which have been compressed to 64 bits each, resulting in about 395-790 bytes per image [¶]. This direct result is competitive with more complex post-encoding strategies based on Fisher and residual vectors that produce about 350-600 bytes per image but using complex processing and requiring special training.[9–11]

We have tested 3 key-point detection methods: SIFT key-point detector,[5] FAST-9 used in ORB[12] and the Harris detector[15] and all variants. The SIFT and Harris key-point detectors provide an excellent uniform coverage of the entire image and thus perfectly satisfy requirement **R2**. The FAST and ORB key-point detector are really fast but produce very clustered appearances of key-points. Therefore, it does not directly fit to **R2**. The robustness of points tested to estimate the probability of correct key-point detection P_D^p as requested by **R3.1** was repartitioned as: "SIFT" 0.84, "Harris" 0.85 and "FAST" 0.9 when the targeted number of extracted key-points was above 500 per images under typical distortions of mobile phone imaging that included projective transform, additive noise with the standard deviation 10, and JPEG compression with quality factor 80. The

[¶]Further compression can be obtained via known aggregations method which are out of scope of this paper.

superior performance of FAST is explained by the redundant clustered appearance of points in the vicinity of corners that leads to a natural redundancy, but with a very poor coverage of the image. Even though this property does not satisfy **R2**, the number of key-points should be restricted for the complexity reasons in view of the corresponding combinatorics of all possible read outs. Therefore, the redundancy of FAST might be of great advantage for the robustness, if properly used. Thus, we will target a relatively small number of resulting robust key-points in the order of $60-70$ for complexity reasons. It should be noted that the key-point detector of SIFT and Harris have no integrated key-point selection option. Therefore, one can use various heuristic selection rules such as the intensity of the gradient, or the scale-space in the key-points for the selection of preferred key-points or just a random sampling. These methods however, do not really work well.

3.2 Key-points filtering

The objective of key-point filtering based on FAST is to satisfy the requirements of **R2** and **R3.1** with an approximate number of resulting key-point in the range of 60-70. Therefore, using the initial key-points produced by FAST-9 with a controllable threshold, we performed spatial graph clustering of the points and used only those clusters where the number of originally detected key-points exceeded 2. These groups of points have proven to be more robust than other detectors and feature selection methods. In this case, one can expect that the selected key-points in the enrolled image will correspond to those in the probe with high probability, given an area of acceptance with a radius of 4 pixels. Finally, our test on the distorted images with projective transforms, additive noise up to the standard deviation 30, followed by gamma correction and JPEG-80 compression, resulted in $P_D^p = 0.94$ which is the highest result among all tested key-point detectors. They are thus the most suited for SketchPrint extraction.

3.3 SketchPrints extraction

Figure 3 shows the generalised extraction of the *sketch* descriptor given two reference key-points a and b. At this stage we also require that the distance between the reference points does not exceed twice the number of later used interpolation points, nor may it be less then half this number. These constraints are also introduced to cope with the projective transforms and lens distortions to map all these transformations into an approximate scaling between two key-points.

Since the sketch is taken on a discrete grid, the coordinates of pixels along the line belonging to the sketch might change depending on the orientation and rasterisation of the image. Therefore, to ensure invariance to this sampling ambiguity as well as to benefit from the redundancy in face of possible noise or compression artefacts, we extract several parallel sketches between two key-points within their 3x3 neighbourhood. Each sketch is then re-scaled to a fixed length L. We have used $L = 128$ for comparison reasons with SIFT which has the same length. Then the estimator is applied to all samples $1 \leq i \leq L$ to produce a resulting sketch. In our implementation the mean and median computed from the extracted traces showed near similar performance.

To achieve invariance to amplitude scaling the above signal is transformed into a zero-mean and unit-variance vector by normalization. We also compared normalization to deploying a gradient which is commonly claimed to be more robust to amplitude modifications. Our tests, however showed that the amplitude normalization in the described setting gave superior performance to a gradient based one.

Finally, the real valued normalized sketch is quantized using k-means or product vector quantization to produce a binary vector of length ℓ. In our basic implementation, we have used $\ell = 64$ bits for comparison reasons with quantized SIFT and GHOC reported in.[9,11]

The number of descriptors K_w per image might be varying from one image to the other. Moreover, some descriptors might not satisfy the requirement (**R1**) and might be too high for storage as in the case of SIFT that produces between 500 (real images) to 5'000 (text docs) and 7'000 (microstructure images) descriptors. For these reasons, we apply SketchPrint filtering.

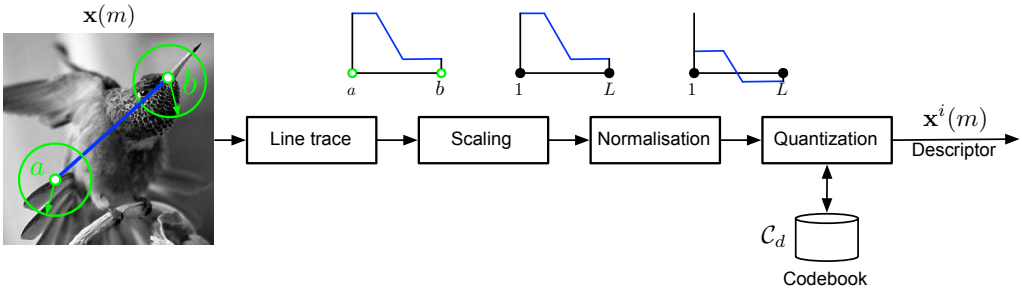

Figure 3: Skecthprint descriptor extraction.

3.4 SkechPrints filtering

To satisfy the requirement (**R1**) and to produce a controllable number of descriptors for efficient memory storage, we apply filtering to the cross-sections extracted at the previous stage.

The goal is to select the most informative cross-sections. To maximize the information content of our descriptors expressed by its entropy, we will assume the local variance to be a suitable parameter for the maximization of entropy under the bounded total variance.

Therefore, the descriptors containing a significant portion of flat regions should be avoided. We will select the SketchPrints with maximum local variances. The SketchPrints that do not satisfy this condition are rejected.

Given the normalized cross-section $\mathbf{x}(m, j)$ for an image m with $1 \leq j \leq K_m$, we compute the local variance in a sliding window of size $W = 5$ using the ML estimation $\sigma^2_{X_i}(m, j)$, $1 \leq i \leq L$. The resulting estimates are sorted to produce order statistics $\sigma^{OS2}_{X_i}(m, j)$, excluding 10 samples at the beginning and end of the cross-section that correspond to the high variability in the vicinity of the key-points that are not informative since they are present in all cross-sections extracted between two key-points. We require the order statistics $\sigma^{OS2}_{X_{25}}(m, j) \geq T$, where $T = 15$ to comply with the informativeness condition.

As a result we end up with J_m descriptors per image m. In case, a fixed number of descriptors J is needed for all images, the descriptors possessing the smallest local variances are filtered out. In addition, very efficient filtering has been achieved based on clustering of similar descriptors that might result from several near-parallel read-outs and keeping only those closest to the "virtual" centroids as the most representative ones.

3.5 SketchPrint descriptor on different content

To demonstrate the universality of the SketchPrint descriptor to simultaneously produce a very informative content representation for images with a different nature, we exemplify several cases for natural images, text/logos and random microstructures (Figure 5). Obviously, such a level of distinguishability can not be achieved by existing local descriptors.

4. EXPERIMENTAL RESULTS

4.1 Descriptor performance evaluation: a perfect synchronization

To have a fair comparison of SketchPrint with state-of-the-art descriptors such as SIFT (the best but slow) and ORB (fast but worser) we investigate its distinguishability and robustness on the UCID database with real 1'338 images[16] which have been distorted with a projective transformation, followed by AWGN and JPEG compression and gamma correction. According to the standard descriptor evaluation procedure, the positions of descriptors from an original image and distorted copy were assumed to be known. Although this situation is artificial for practical usage since the geometrical correspondence between the descriptors is unknown and the identification is done purely based on the descriptor similarity, nevertheless it gives an idea about the robustness and discriminative power of descriptors. Figure 6 shows the ROC curves for the 3 descriptors. SketchPrint descriptor was tested in two versions: (a) intensity normalization and (b) gradient along the cross-section. SketchPrint remarkably outperforms other descriptors on natural images as it retains by far the most original

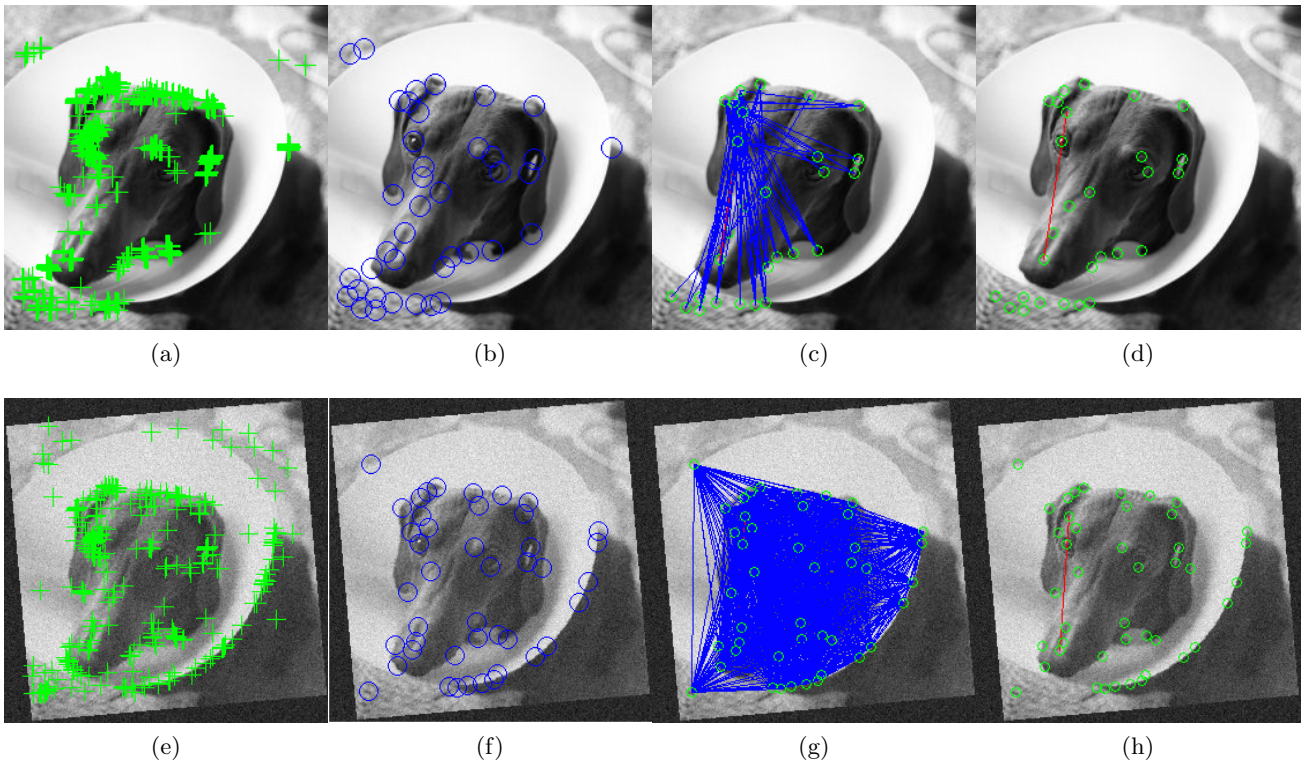

Figure 4: SketchPrint descriptor computation at the main stages for the original image (upper line) and distorted image (bottom line): (a),(e) original and distorted images with the FAST key-points detection, (b),(f) key-points filtering, (c),(g) SketchPrints extraction and (d),(h) SketchPrints filtering (example of a match).

image information. Its performance on text documents and microstructures is even more impressive. The SketchPrint based on the normalized intensity also outperforms the gradient based version. Finally, it is worth mentioning that SketchPrint is very distinctive as is proven by the very low probability of P_F. Practically, it means that only several descriptors might suffice for a unique image representation. SketchPrint robustness expressed by P_M is also considerably lower that this for SIFT and ORB.

4.2 Descriptor performance evaluation: a geometrically blind matching

The above test is artificial in the sense of assuming a known geometrical correspondence between descriptors. In practice it is not the case since the descriptors are used for the approximate but fast estimation of a possible list of images similar to the probe with similar appearances of descriptors. That is why it is very important to evaluate the descriptor "blindly", i.e., without any prior information about the geometric correspondence. In fact, this regime exactly corresponds to the basic design of BoF as considered in Section 2. In this case, the identification system retrieves a set of images or preferably just one for which the maximum number of descriptors extracted from the probe are as close as possible to the descriptors of the most similar image(s) in the Euclidean or Hamming space. To accelerate the search, the comparison is not done directly which would be prohibitively complex but via a reference set of codewords that are the most representative for a given dataset. Generally, it requires a lot of optimization and many parameters to be optimally tuned of a given dataset.

Therefore, to demonstrate the core performance of the proposed descriptor, we will just show the distribution of distances between the probe descriptors (all extracted) and a restricted number of descriptors enrolled from the original image. As a probe we will use the distorted original image and just another image from the same application that represents the "closest" semantic challenge. As applications, we consider the identification of text documents (without images), packages (a combination of text and graphics) and microstructure images (random correlated noise) as shown in Figure 7.

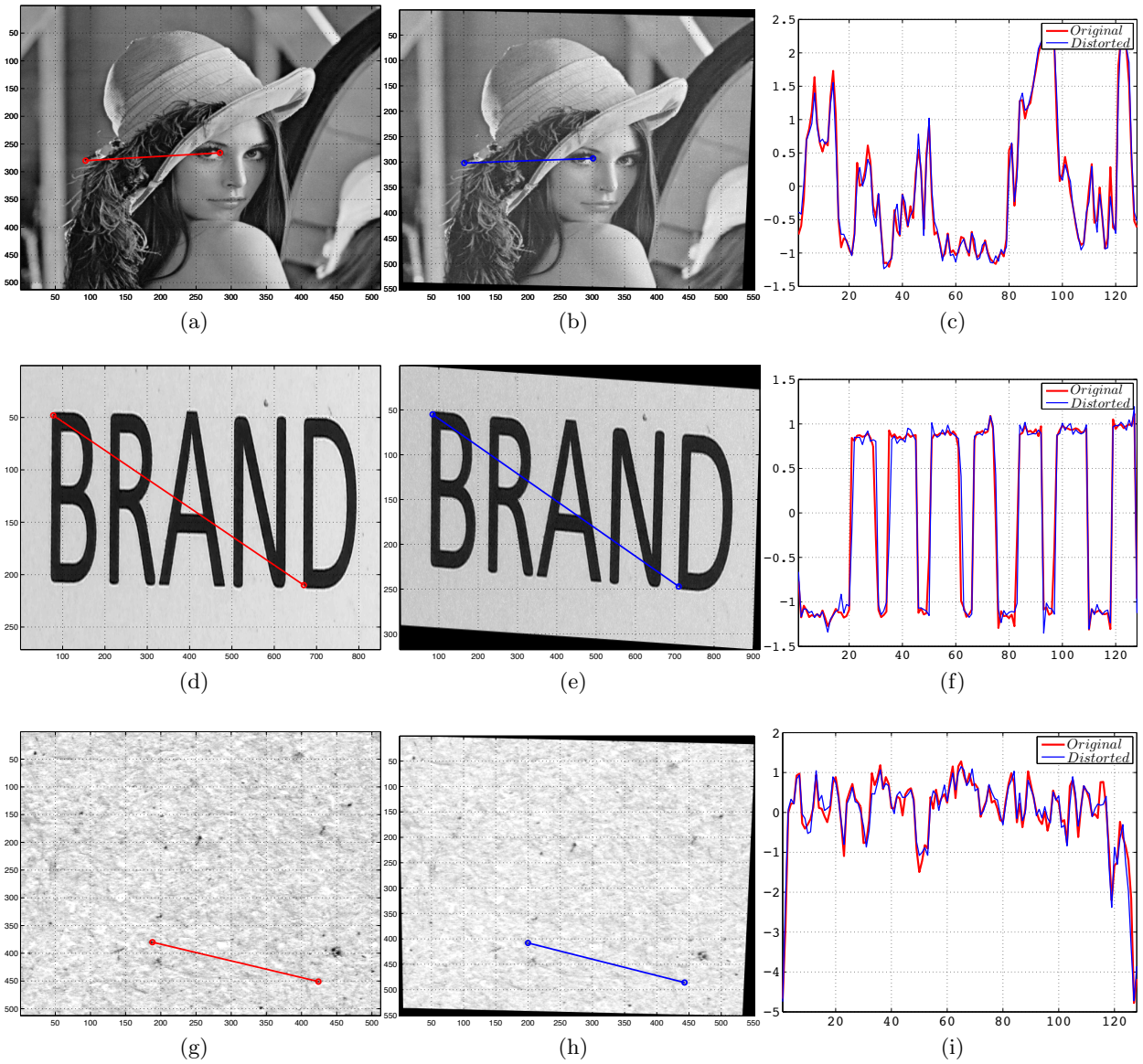

Figure 5: Examples of SketchPrint descriptors extracted from the original (a),(d),(g) and distorted images (b),(e),(h) and their comparison (c),(f),(i).

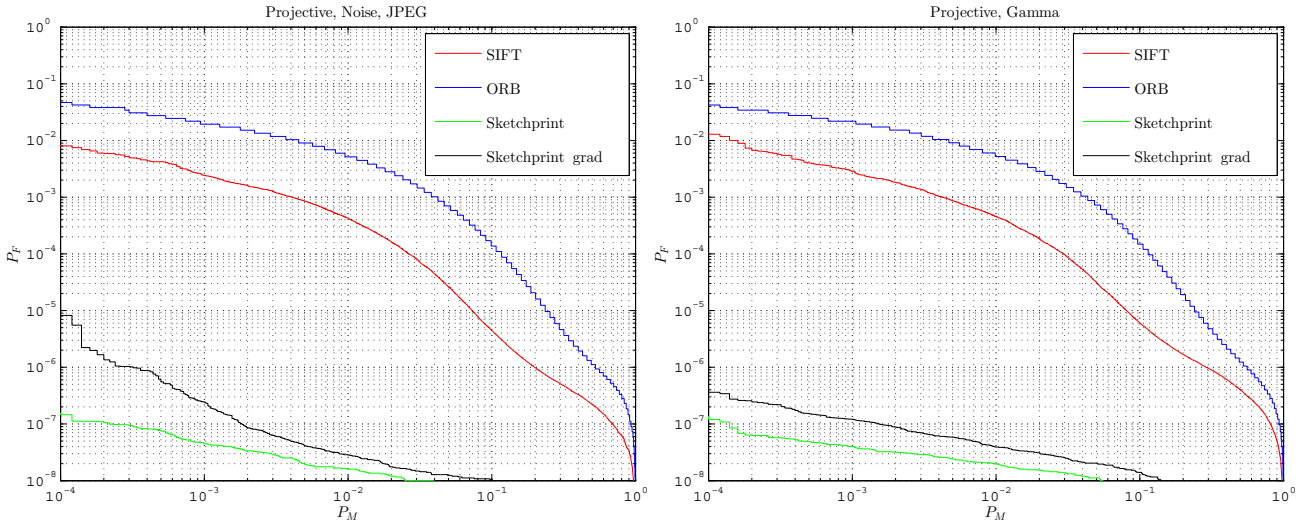

Figure 6: Comparison of ROCs for SIFT, ORB and SketchPrint descriptors, using a priori geometrical information to exclude the impact of the key-point detectors under: (a) projective transform, AWGN with noise standard deviation 10 and JPEG compression QF=80 and (b) projective transform and gamma correction.

In our tests, we have enrolled a maximum of a 100 SketchPrints, 100 and 1'000 SIFT descriptors chosen based on the maximum gradient magnitude in each point[‖]. The probe contained an unconstrained number of extracted descriptors. Since a real practical system does not have any information about the geometrical correspondence between the descriptors, all descriptors extracted from the probe are tested exhaustively against the enrolled descriptors using the Euclidean distance. If one had the geometrical descriptors' correspondence, one would plot the corresponding histograms for inliers and outliers as in the previous section. Therefore, we investigated the statistics of ordered distances for the corresponding pairs of images for different data sets.

As a test for a base performance and correlation amongst descriptors from different images of the same category, the order statistics of distances between the enrolled 100 SketchPrints, 100 SIFTs and 1'000 SIFTs are shown in Figure 8. It clearly shows that enrolled SketchPrint descriptors are unique, i.e., the smallest distance between two different images is higher that the distance with the order 100 for SketchPrint for the same image (becides image (d) where less descriptors have been enrolled). For SIFT descriptor this condition is not always satisfied especially for the case of 1'000 descriptors.

The final results where a distorted probe is matched against enrolled examples can be seen in Figure 9. If the probe corresponds to the correct image, the order statistics of distances should slowly grow from the smallest Euclidean distance which corresponds to the best attained match. In contrast, if the probe is taken from the wrong image, the distances are very large. If the descriptors are unique and robust and the probe's descriptors do not contain a lot of redundant descriptors, the flat slope of the matched distance order statistics should extend up to the number of enrolled descriptors in the ideal case. After this point, the order statistics should increase rapidly to the level of the non-matched ones. At the same time, the largest "matched" distance in the list of order statistics should not exceed the minimum one for the non-matched images. This condition guarantees that the images are clearly distinguishable. The results for SketchPrint clearly indicate that one can achieve a performance competitive to 1000 SIFT and superior to 100 SIFT performance. More particularly, the results for the tested contents indicate the distinguishability for: 100 SketchPrint - 100%, 90%, and 90%; 100 SIFT - 40%, 75%, and 45%; 1'000 SIFT - 75%, 30%, and 75% for text documents, packages and microstructures, respectively.

[‖]The original algorithm doesn't offer this option, therefore, it is our own implementation. Otherwise, SIFT produces from 2'000 to 5'000 descriptors per image which is not fair for our comparison.

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 $\mathrm{E}\{.\}$ 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.

(a)

Gaussian Wiretap Channel with Collaborative Wiretappers

Svyatoslav Voloshynovskiy, Taras Holotyak, Ivan Prudyus

***Abstract*—In this paper, we consider the compound wiretap channel with collaborative wiretappers. Contrarity to [1], where the wiretappers act independently, we analyze the setup when the wiretappers form a coalition, which shares the observed data within the coalition. The goal of coalition consists in learning the secret message communicated to the legitimate user by sharing their observations to produce the best possible estimate thus benefiting from available redundancy. We analyze the secrecy capacity for the Gaussian data and correlated observations. As the result, we provide the estimate of the impact of coalition of wiretappers on reduction of secrecy capacity.**

***Index Terms*—Wiretap channel, secrecy capacity, collaborative wiretappers.**

I. Introduction

The wiretap channel model of Wyner appeared as a noisy counterpart of noiseless secure information transmission problem of Shannon [2]. In Wyner wiretap model [3], essentially linked to the secure extension of broadcast channel [4], the advantage of legitimate information receiver in the level of noise in data is exploited over the opponent who observes the same transmission in more noisy environment. Several cases of discrete memoryless channels (DMC) including Gaussian one were studied in [5] and it was demonstrated that the secrecy capacity is equal to the difference between the capacity of the main channel connecting the sender and the legitimate user and the wiretap channel. Csiszar and Korner in [6] extended this model and characterized the capacity of DMC under security constraints. Liang et al [1] further extended these results to compound channel where both legitimate and wiretap channels might take a number of states. The most related considered case is based on the semi deterministic compound wiretap channel where the secrecy capacity is found. This case includes the legitimate receiver and the the group of wiretappers that observe independently the transmission. The secrecy capacity is determined by the worst channel among Such a form of collaborative coalition is able to benefit from the redundant noisy observations and can produce better estimate of the secret message in comparison to the individual estimates analyzed in [1]. If the number of wiretappers in the coalition is sufficient, the coalition can disclose the secret message. It also means that the secrecy rate approaches zero and no secure transmission is possible anymore. At the same time, the distortions in the wiretapper channels might be correlated and it is important to establish the theoretical limits of this coalition. Therefore, the goal of this paper is to establish the secrecy rate under collaborative coalition of wiretappers and to investigate the impact of dependent observations for the Gaussian and binary observation models.

II. Problem formulation

We consider the following wiretap channel model.

Definition 1 (*degraded wiretap channel*): The degraded wiretap channel consists of input alphabet $\mathcal{X}^N$, the output alphabet of legitimate user $\mathcal{Y}^N$, J channel outputs of wiretappers $\mathcal{V}^N$ and the corresponding transition probabilities for the legitimate user, assumed to be memoryless, $p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p(y_i|x_i)$ and the wiretappers $p(\mathbf{v}_j|\mathbf{y}) = \prod_{i=1}^{N} p(v_{ij}|y_i)$, where $\mathbf{x} \in \mathcal{X}^N$ is the channel input generated by the encoder, $\mathbf{y} \in \mathcal{Y}^N$ is the channel output available for the legitimate user, $\mathbf{v}_j \in \mathcal{V}_j^N$ is the wiretap channel output available for the j wiretapper.

Definition 2 (*code construction*): The $(2^{NR}, N)$ code for the wiretap channel consists of:

- a *message set* $\mathcal{W} = \{1, 2, \cdots, 2^{NR}\}$ with the message M uniformly distributed over $\mathcal{W}$;
- an *encoder* ϕ: $\mathcal{W} \mapsto \mathcal{X}^N$;
- a *decoder* σ: $\mathcal{Y}^N \mapsto \mathcal{W}$

(b)

(c)

(d)

(e)

(f)

Figure 7: Examples of the used text documents, packages and random microstructure images.

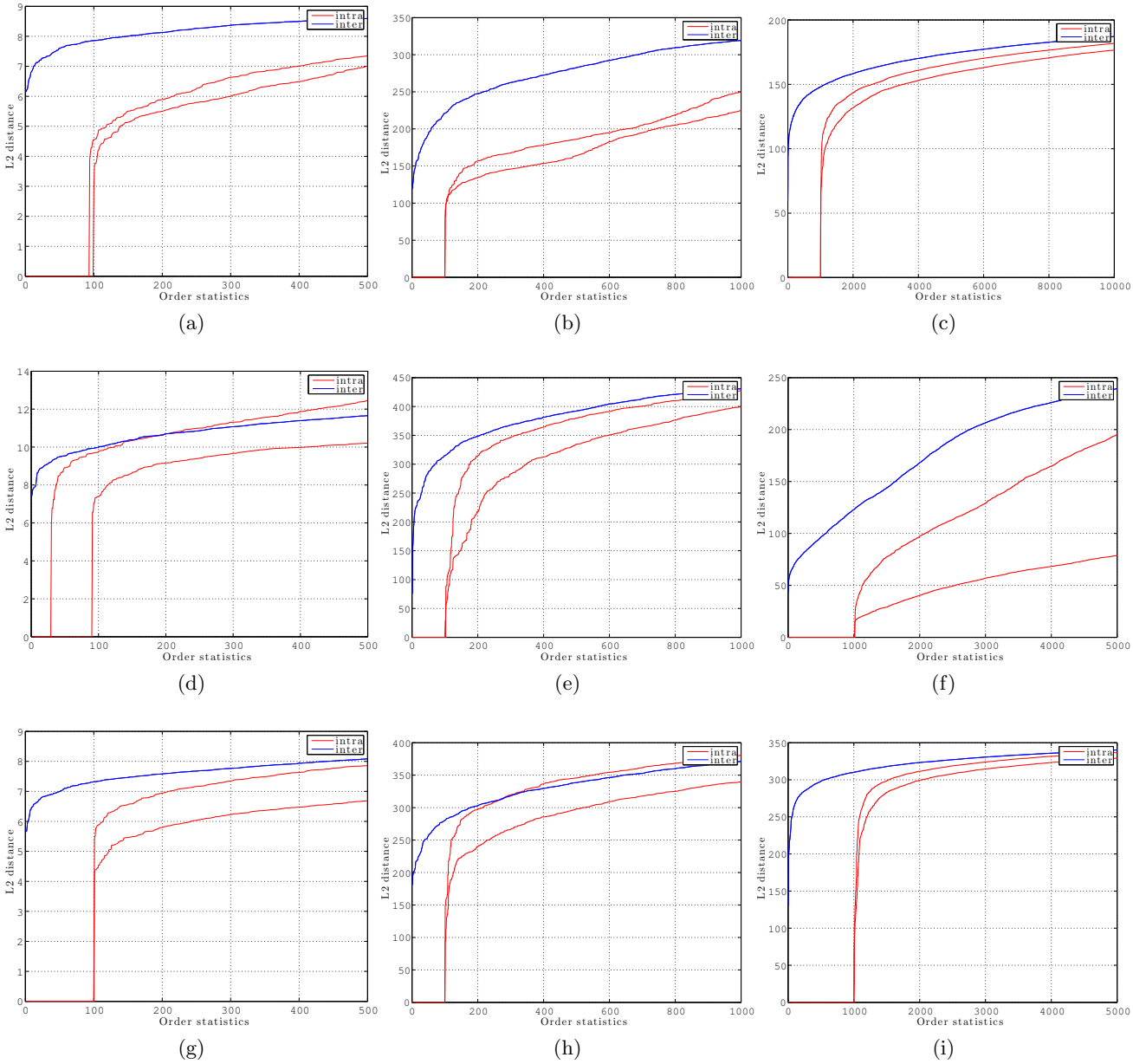

Figure 8: 100 enrolled SketchPrints (a), (d), (g), 100 enrolled SIFTs (b), (e), (h) and 1'000 enrolled SIFTs (c), (f), (i) performance on text documents (a-c), packages (d-f) and random microstructure images (g-i) where the probe and the enrolled item were identical (self-assestment).

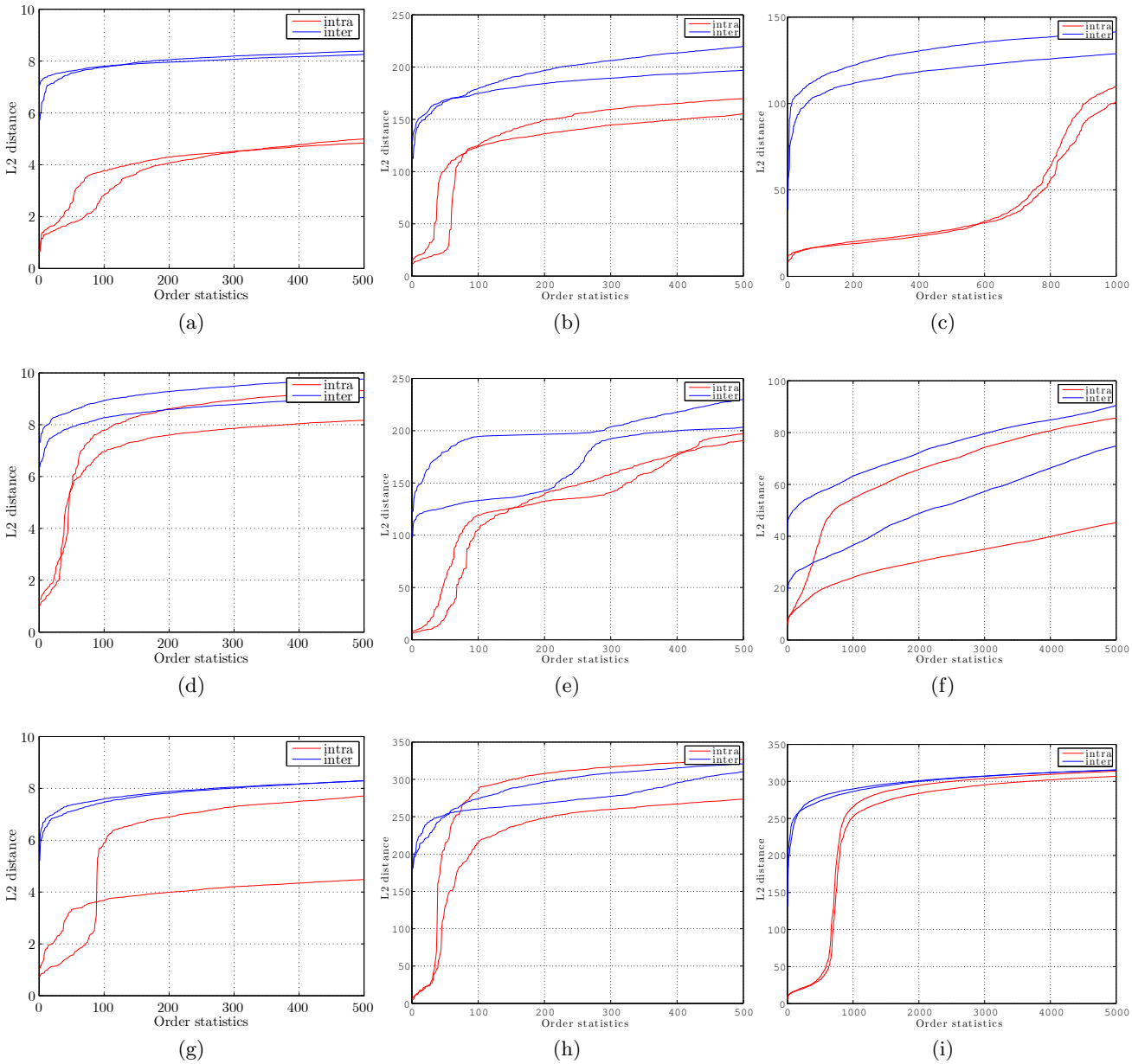

Figure 9: 100 enrolled SketchPrints (a), (d), (g), 100 enrolled SIFTs (b), (e), (h) and 1'000 enrolled SIFTs (c), (f), (i) performance on text documents (a-c), packages (d-f) and random microstructure images (g-i) on the original and a distorted copy.

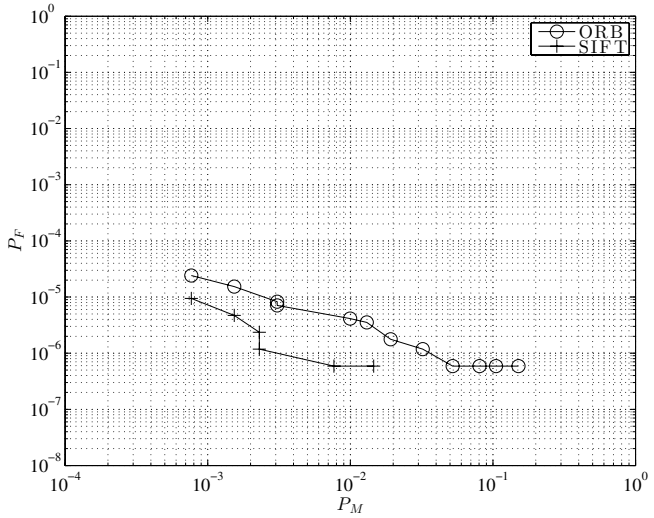

Figure 10: Identification performance for SketchPrint (errorless), SIFT and ORB on the UCID database.

4.3 Identification performance evaluation: no geometrical matching

Finally, to validate SketchPrint in full identification mode, we used the UCID database of real 1'338 images[16] which have been distorted with a projective transformation, followed by AWGN and JPEG compression. In an extreme case, this database might be stored on a mobile device. Therefore, we stipulated that the image would be represented just by 32 SketchPrint, SIFT and ORB descriptors per image. The identification test was based on the above ordered statistics with the decision rule based on the 3rd ordered distance. Although the database is relatively small, it was our intention to see the system performance when no BoF compression is used which only decreases the identification performance and a very limited amount of descriptors. The SketchPrints have been of length 128 and quantized to 8 bits per sample to be comparable with the SIFT descriptors while the ORB descriptors had 256 bits. The systems based on SketchPrint identified all images errorless with different distortions. The ROC plot for SIFT and ORB is shown in Figure 10.

5. CONCLUSIONS

In this paper, we present a new type of image representation based on the SketchPrint descriptor. In contrast to most of the popular local descriptors like SIFT, SketchPrint extracts non-local information between any two reference key points and normalizes it. Being conceptually simple and computationally efficient, SketchPrint provides a very unique and robust image representation which can not attained by any tested local descriptor. Finally, only 35 descriptors suffice for errorless image identification in the UCID database, without any post extra geometrical matching and re-ranking. We investigated the upper limits of performance when the descriptors are stored in an uncompressed form and the system is performing an exhaustive identification. In the future work, we will investigate the optimal form of SketchPrint compression and fast indication based on proper indexing.

Acknowledgment

This paper was partially supported by SNF projects 200020-146379.

REFERENCES

[1] Rogers, E., Rodriguez, T. F., Lord, J., and Alattar, A. M., "Performance evaluation of digimarc discover on google glass," in [*Media Watermarking, Security, and Forensics 2015*], (January 2015).

[2] Voloshynovskiy, S., Diephuis, M., Beekhof, F., Koval, O., and Keel, B., "Towards reproducible results in authentication based on physical non-cloneable functions: The forensic authentication microstructure optical set (famos)," in [*Proceedings of IEEE International Workshop on Information Forensics and Security*], (December 2–5 2012).

[3] Diephuis, M. and Voloshynovskiy, S., "Physical object identification based on famos microstructure fingerprinting: comparison of templates versus invariant features," in [*Proceedings of 8th International Symposium on Image and Signal Processing and Analysis (ISPA 2013)*], (September 4–6 2013).

[4] Diephuis, M., Voloshynovskiy, S., Holotyak, T., Standardo, N., and Keel, B., "A framework for fast and secure packaging identification on mobile phones," in [*Proceedings of SPIE Photonics West, Electronic Imaging, Media Forensics and Security V*], (January, 23 2014).

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

[6] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A., "Object retrieval with large vocabularies and fast spatial matching," in [*IEEE Conference on Computer Vision and Pattern Recognition*], (2007).

[7] Ondřej Chum, Michal Perďoch, J. M., "Geometric min-hashing: Finding a (thick) needle in a haystack," in [*IEEE Computer Society Conference on Computer Vision and Pattern Recognition*], 17–24 (June 2009).

[8] Lazebnik, S., Schmid, C., and Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in [*Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*], **2**, 2169–2178, IEEE (2006).

[9] Jégou, H., Douze, M., Schmid, C., and Pérez, P., "Aggregating local descriptors into a compact image representation," in [*Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*], 3304–3311, IEEE (2010).

[10] Perronnin, F., Sánchez, J., and Mensink, T., "Improving the fisher kernel for large-scale image classification," in [*Computer Vision–ECCV 2010*], 143–156, Springer Berlin Heidelberg (2010).

[11] Girod, B., Chandrasekhar, V., Chen, D. M., Cheung, N.-M., Grzeszczuk, R., Reznik, Y., Takacs, G., Tsai, S. S., and Vedantham, R., "Mobile visual search," *Signal Processing Magazine, IEEE* **28**(4), 61–76 (2011).

[12] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., "Orb: an efficient alternative to sift or surf," in [*Computer Vision (ICCV), 2011 IEEE International Conference on*], 2564–2571, IEEE (2011).

[13] Calonder, M., Lepetit, V., Strecha, C., and Fua, P., "Brief: Binary robust independent elementary features," in [*Computer Vision–ECCV 2010*], 778–792, Springer (2010).

[14] Bay, H., Tuytelaars, T., and Gool, L. V., "Surf: Speeded up robust features," in [*In ECCV*], 404–417 (2006).

[15] Harris, C. and Stephens, M., "A combined corner and edge detector," in [*Proceedings of the 4th Alvey Vision Conference*], 147–151 (1988).

[16] Schaefer, G. and Stich, M., "Ucid-an uncompressed colour image database," *Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia* **5307**, 472–480 (2004).