Content identification: binary content fingerprinting versus binary content encoding

Sohrab Ferdowsi, Svyatoslav Voloshynovskiy, Dimche Kostadinov

University of Geneva, Battle Bât. A, 7 route de Drize, 1227 Carouge, Switzerland;

ABSTRACT

In this work, we address the problem of content identification. We consider content identification as a special case of multiclass classification. The conventional approach towards identification is based on content fingerprinting where a short binary content description known as a fingerprint is extracted from the content. We propose an alternative solution based on elements of machine learning theory and digital communications. Similar to binary content fingerprinting, binary content representation is generated based on a set of trained binary classifiers. We consider several training/encoding strategies and demonstrate that the proposed system can achieve the upper theoretical performance limits of content identification. The experimental results were carried out both on a synthetic dataset with different parameters and the FAMOS dataset of microstructures from consumer packages.

Keywords: Content Identification, Content Fingerprinting, Multiclass Classification, ECOC.

Notation

Scalar random variables are denoted with capital letters X while the vector version as $\mathbf{X}$ and their realisations as x and $\mathbf{x}$, respectively, i.e, $\mathbf{x} = \{x(1), x(2), \cdots, x(N)\}$. $X \sim p(X)$ is used to show that a random variable follows the distribution $p_X(x)$. We denote the identity matrix of size N as $\mathbb{I}_N$.

1. INTRODUCTION

Content identification has emerged in many applications. In digital media, for example, we face the problem of image, audio or video identification for copyright protection, tracking and tracing. In physical world security, physical uncloneable functions are used to provide reliable and secure identification and authentication.

Multimedia data, as well as features acquired from physical objects or humans, normally lie in very high-dimensional spaces. Moreover, in many applications, there are also the privacy preservation and security concerns to be considered. To address these issues, a *content fingerprint*, a.k.a. *robust perceptual hash* is proposed as a solution. Content fingerprinting generates a short, robust, binary content representation that can efficiently be shared and securely saved between different parties. It is mostly implemented as a combination of dimensionality reduction transform and quantisation which is often implemented in a form of binary quantizers. The resulting binary fingerprints can be easily identified using fast methods such as Hamming distance or advanced nearest neighbour search strategies. Being simple and tractable, the content fingerprinting is an information lossy transform. It is also known that it leads to the loss in performance in terms of both achievable rate and identification accuracy.[1]

Moreover, in many applications there is more than one representation of the same content object or human biometrics acquired under different conditions or time instances available in the database. Naturally, such a redundant representation should be superficial for identification in varying and noisy channels. Content fingerprinting, however, does not take into account such a redundancy. At the same time, the methods developed in modern machine learning are explicitly based on learning where all prior information could be used to design efficient classifiers.

From another viewpoint, the content identification problem can also be treated as a multiclass classification problem.[2,3] In this framework, each object in the database is considered as an instance belonging to a different

Further author information: (Send correspondence to Sviatoslav Voloshynovskiy)
E-mail: svolos@unige.ch, Telephone: 41 22 3790158, http://sip.unige.ch

class whose boundaries should be trained based on the data from the items in the database. Then a query is identified to belong to an item in the database by classifying it into one of the previously trained classes. An important challenge in this method, given the potentially huge size of the database, is how to formulate the multiclass classification problem based on the common binary classification methods so that the complexity be tractable. We advise an existing approach known as Error Correcting Output Codes (ECOC) which is linked with coding theory. We propose, however, a novel method to treat this problem efficiently.

We investigate the methods first on a synthetic dataset of randomly generated multivariate Gaussians as the authentic objects and added different levels of Gaussian noise to each of the items as the query data to be identified. We then run the same experiments on the Forensic Authentication Microstructure Optical Set (FAMOS)[4] that contains 5000 physical objects acquired by two different cameras each three times.

The paper is organized as follows: In section 2 we state the problem formulation and explain a natural approach to solve it in the real domain. In section 3 we consider the fingerprinting approach in the binary domain. We have a detour from the problem of identification in section 4 to introduce the problem of multiclass classification and a coding theory based approach to solve it. We then come back to the identification problem and propose a novel approach in section 5 to address it. Experimental results are explained in section 6 and we conclude the paper in section 7.

2. PROBLEM FORMULATION

In this section we introduce the general problem of identification when the data is in the real domain. Due to its complexity, memory storage and privacy preservation issues, this method is not used in practice. However, the materials in this section provide a good understanding of the whole problem and indicate the upper bounds on the performance for other binary methods.

2.1 Problem Formulation in Real Domain

The data owner provides $|\mathcal{M}|$ entries to the database consisting of $\mathbf{x}(m)$'s, where $1 \leq m \leq |\mathcal{M}|$. Then depending on the application, a stage of invariant feature extraction is applied to get $\mathbf{v}_x(m) \in \mathbb{R}^{\mathbb{N}}$ as shown in Figure 1. The idea is to have geometrical robustness against different de-synchornization variations between the enrollment and identification. The geometrical invariance transform $\mathbb{F}$ maps the input data $\mathbf{x}(m)$ into the feature vector $\mathbf{v}_x(m)$ using either: (a) explicit synchronisation based on some marks of printed symbologies, or (b) invariant features previously learned via the codebook $\mathcal{D}$. The invariant features might include the popular local invariance descriptors such as SIFT or the trained patches.

Then, in order to reduce complexity and to map the data to the domain with the expected statistical properties, we perform an optional stage of dimensionality reduction from N to $L \leq N$ through random projections with a matrix $\mathbb{W} \in \mathbb{R}^{L \times N}$:

$$\mathbf{r}_x(m) = \mathbb{W}\mathbf{v}_x(m), \tag{1}$$

where $\mathbb{W} = (\mathbf{w}_1; \mathbf{w}_2; ...; \mathbf{w}_L)^T$ consists of a set of projection basis vectors $\mathbf{w}_i \in \mathbb{R}^N$ with $1 \leq i \leq L$. In our study the elements of $\mathbb{W}$ are chosen as independent and identically distributed realisations of a Gaussian random variable $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$. Consequently, we could consider $\mathbb{W}$ to be an almost orthoprojector where $\mathbb{W}\mathbb{W}^T \approx \mathbb{I}_L$. Other than complexity reduction, random projections, while proved to preserve the distance between the data, under certain conditions, also decorrelates the data samples.[5] As a result, it leads to independence or weak dependence between the samples depending on the statistics of vectors $\mathbf{v}_x$. After performing random projections, the resulting features are stored as $\mathbf{r}_x(m)$'s in database $\mathcal{B}_r$.

At the identification phase, the data user provides query data $\mathbf{y}$ which is to be identified as one of the $|\mathcal{M}|$ objects in the database. It is considered to be probabilistically related to the m'th item of the database $\mathbf{x}(m)$ through $p(\mathbf{y}|\mathbf{x}(m))$. From the given $\mathbf{y}$, we generate $\mathbf{v}_y$ and $\mathbf{r}_y$ through the same procedure as at the enrolment stage. Then we identify the given query as one of the items in the database based on a similarity measure between $\mathbf{r}_y$ and $\mathbf{r}_x(m)$'s. Figure 1 sketches these steps. The identification problem is essentially an $|\mathcal{M}|$-ary classification implemented in the decoder $\mathbb{D}$ with an optional rejection option that will not be considered in this paper.[*]

[*] For more information about the impact of the rejection option, please see[6]

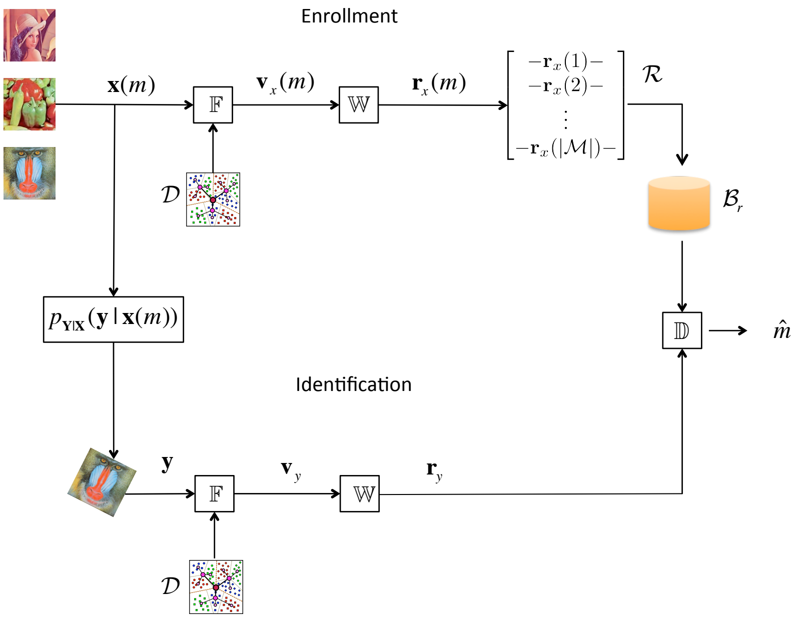


Figure 1. General identification scenario in the real domain. $\mathbb{F}$ provides geometrical invariance, $\mathbb{W}$ stands for random projections and $\mathbb{D}$ is the minimum Euclidean distance or bounded distance decoder.

We assume a Gaussian distribution for our data, i.e, $\mathbf{r}_x(m)$'s are *i.i.d.* realisations of $\mathbf{R}_x \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbb{I}_L)$. The $\mathbf{r}_y$'s are considered as noisy versions of $\mathbf{r}_x$'s, i.e., $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_z$ and $\mathbf{R}_z \sim \mathcal{N}(\mathbf{0}, \sigma_z^2 \mathbb{I}_L)$. Based on this formulation we can analyse the performance of this system based on the probability of correct identification and achievable identification rate. The assumption on the *i.i.d.* Gaussian statistics of randomly projected data in the content fingerprinting context is based on the proof provided in[5]

It is also shown that the $|\mathcal{M}|$-ary classification can be reduced to maximum likelihood decoding under the equally likely classes, i.e., $\Pr\{M\} = 1/|\mathcal{M}|$. Therefore, the identity of the query is decided as:

$$\hat{m} = \underset{1 \leq m \leq |\mathcal{M}|}{\operatorname{argmax}} \; p(\mathbf{r}_y | \mathbf{r}_x(m)). \tag{2}$$

Since we are in the real domain and the noise is Gaussian, the problem reduces to minimum Euclidean distance decoding:

$$\hat{m} = \underset{1 \leq m \leq |\mathcal{M}|}{\operatorname{argmin}} \; ||\mathbf{r}_y - \mathbf{r}_x(m)||_2, \tag{3}$$

where $||.||_2$ is the Euclidean distance between two vectors.

2.2 Performance Analysis

We define the probability of correct identification when we receive $\mathbf{r}_y$ while the true object was the m'th item in the database as:

$$P_{ci} = \frac{1}{|\mathcal{M}|} \sum_{1 \leq m \leq |\mathcal{M}|} \Pr\left\{\hat{M} = M | M = m\right\}, \tag{4}$$

which yields:

$$P_{ci} = \Pr\left\{||\mathbf{R}_y - \mathbf{r}_x(m)||_2 < \min_{1 \leq m' \neq m \leq |\mathcal{M}|} ||\mathbf{R}_y - \mathbf{r}_x(m')||_2\right\}. \tag{5}$$

This problem could also be studied in an information-theoretic framework. Due to the Asymptotic Equipartition Property, when the data are independent, the maximum number of items that could be distinguished with vanishing probability of error $P_{ci} \to 0$ as the data dimension L grows large is bounded as:[7]

$$|\mathcal{M}| \leq 2^{LI(R_x;R_y)} \leq 2^{LC_{GC}}, \tag{6}$$

where $I(\cdot;\cdot)$ is the mutual information operator and C_{GC} is the Gaussian channel identification capacity which is equivalent to the amount of mutual information between R_x and R_y. The identification channel capacity for the considered Gaussian setup equals:

$$C_{GC} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_x^2}{\sigma_z^2}\right). \tag{7}$$

3. BINARY CONTENT FINGERPRINTING

To cope with the privacy protection, search complexity and memory storage issues, identification systems, rather than working in the original real domain of the data, are supposed to operate in the binary domain. Therefore, we are interested to consider short binary representations of the items in the database. A natural idea would be to use Vector Quantisation (VQ) $\mathbb{Q}$ with appropriate robust labelling like Grey codes; here we simply binarise the data based on a product of scalar quantisers to get $\mathbf{f}_x(m)$'s from $\mathbf{r}_x(m)$'s and $\mathbf{f}_y$ from $\mathbf{r}_y$ as:

$$\mathbf{f}_x(m) = \{sign(r_{x,1}(m)), \cdots, sign(\mathbf{r}_{x,L}(m))\}, \tag{8}$$

$$\mathbf{f}_y = \{sign(r_{y,1}), \cdots, sign(\mathbf{r}_{y,L})\}, \tag{9}$$

where $\mathbf{f}_x(m)$ and $\mathbf{f}_y(m) \in \{0,1\}^L$ and $sign(\alpha) = 1$, if it is nonnegative and 0, otherwise. We then store these fingerprints in a database $\mathcal{B}_f$ of binary fingerprints. Figure 2 illustrates these steps.

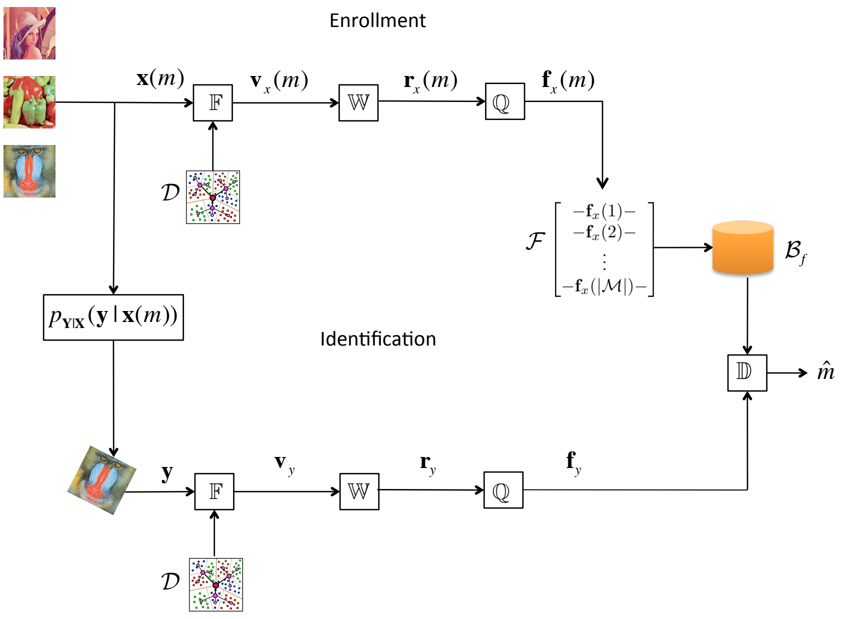


Figure 2. Identification based on fingerprinting. $\mathbb{Q}$ is the quantization function and $\mathbb{D}$ is the minimum Hamming distance decoding.

The identification decision is based on finding a member of $\mathcal{B}_f$ which is closest to the query fingerprint $\mathbf{f}_y$. Equivalently, it could be considered as a decoding step in the communication framework.

Based on the Gaussianity assumptions in the previous section, the mismatch between the data owner fingerprint $\mathbf{f}_x(m)$ and the query $\mathbf{f}_y$ can be modelled as a memoryless Binary Symmetric Channel (BSC). It is shown in[8] that the probability of bit flipping of this BSC is equivalent to $P_b = \frac{1}{\pi} \arccos \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_z^2} \right)$.

In our scenario, the maximum likelihood decoding rule reduces to a minimum Hamming distance rule:[8]

$$\hat{m} = \underset{1 \leq m \leq |\mathcal{M}|}{\operatorname{argmin}} \; d^H(\mathbf{f}_y, \mathbf{f}_x(m)), \tag{10}$$

where $d^H(.,.)$ denotes the Hamming distance operator.

3.1 Performance Analysis

Following the above assumptions, the probability of correct identification in this case is calculated as:

$$P_{ci} = \Pr \left\{ d_H(\mathbf{F}_y, \mathbf{f}_x(m)) < \min_{1 \leq m' \neq m \leq |\mathcal{M}|} d_H(\mathbf{f}_x(m'), \mathbf{F}_y) \right\}, \tag{11}$$

which reduces to:[5]

$$P_{ci} = \sum_{d=0}^{L} \left(\left[\sum_{t=d}^{L} (\frac{1}{2})^L \right]^{|\mathcal{M}|-1} \binom{L}{d} P_b^d (1 - P_b)^{L-d} \right). \tag{12}$$

The maximum number of distinguishable items for large values of L is bounded as:

$$M \leq 2^{LI(F_x;F_y)} \leq 2^{LC_{BSC}} = 2^{L(1-H_2(P_b))}, \tag{13}$$

where $H_2(.)$ is the binary entropy function and P_b is the probability of a bit flipping in the equivalent Binary Symmetric Channel as defined above and the BSC identification capacity is $C_{BSC} = I(F_x; F_y)$.

4. MULTICLASS CLASSIFICATION

In this section we define the problem of multiclass classification in the machine learning formulation. Then we explain an important approach towards solving this problem known as Error Correcting Output Codes (ECOC).[9] The reason behind introducing this subject is that there is a natural analogy between this problem and our identification case. The outputs of binary classifiers forming multiclass labels can be considered as the corresponding bits of fingerprints. At the same time, the fundamental difference between the content fingerprinting and the ECOC multiclass classification consists in the way how the bits are extracted. In the content fingerprinting, the bits are obtained from the random projections that span the space in a random way without any information about the classes. On the contrary, in the ECOC formulation, the projections are trained in the optimal way taking into account all available information about the codeword(s) representing each class. In fact, we will demonstrate that the optimal ECOC training and basis selection repeat the configuration of Voronoi regions that separate classes. As a consequence, we will show that the ECOC represents the ML decoding rule implemented via a set of binary classifiers and the content fingerprinting can only asymptotically approach the performance of ECOC.

In the general case, we are given $|\mathcal{M}| \times p$ training examples, $\mathbf{x}(i) \in \mathbb{R}^N$ as the features and also their labels $g(\mathbf{x}(i))$'s which are generated through an unknown mapping function $g : \mathbb{R}^N \longmapsto \{1, 2, ..., |\mathcal{M}|\}$ that we want to approximate based on the given labeled training examples. The labels, unlike the common cases in machine learning, belong to a set of $|\mathcal{M}| \geq 2$ members rather than only two classes. Because most of the existing classification algorithms are naturally designed for $|\mathcal{M}| = 2$ classes, to generalise them for multiclass cases usually requires to consider the problem as several binary classification problems. Among the existing methods there is the famous *one-vs-all* approach which considers each of the classes in each binary classification problem to be classified against the rest. This requires to run the binary classification algorithm $|\mathcal{M}| - 1$ times. Another method would be to consider each pair of classes together. This method, known as *one-vs-one* approach, requires $\frac{|\mathcal{M}|(|\mathcal{M}|-1)}{2}$ binary classifications to capture all the combinations between the data.

These methods while could work good for small $|\mathcal{M}|$, will seriously fail with the increase of $|\mathcal{M}|$, as the number of binary classifications to be carried out would be intractable, i.e., linear or quadratic in $|\mathcal{M}|$ for the one-vs-all and one-vs-one, respectively. The fact that in order to uniquely define $|\mathcal{M}|$ objects would naturally require $\log_2 |\mathcal{M}|$ binary descriptors has motivated an extensive research in the field. A significant achievement was due to Dietterich[9] who made an analogy between the problem of classification and communication channel and thus suggested to consider it as a channel coding problem. This work brought about a significant amount of research that studied the problem by considering different coding strategies.

4.1 Error Correcting Output Codes (ECOC)

The main idea behind the ECOC approach to solve the above problem is to consider it as a communication system where the identity of the correct output class for a given query is being transmitted over a hypothetical channel which, due to the imperfections of the training data, the errors in the learning process and non-ideal choice of features is considered to be noisy.[9] Therefore, it would make sense to try to encode the classes using error correcting codes and transmit each of the bits through the channel, i.e., to run the learning algorithm, so that we would be able to cope with the errors in each individual binary classifier.

Concretely, we assign randomly to each of the $|\mathcal{M}|$ classes a row of a coding matrix $\mathcal{C}_{(|\mathcal{M}|\times l)}$. Then we run a binary learning algorithm like SVM on all the training samples for every column of $\mathcal{C}$ so that we will have l mappings from $\mathbb{R}^N$, the data space to the one dimensional binary space $\{0, 1\}$, or equivalently, one mapping rule from $\mathbb{R}^N$ to the l-dimensional binary space $\{0, 1\}^l$.

Given a new unlabeled example $\mathbf{y}$, we map it from $\mathbb{R}^N$ to the l-dimensional binary space through the same mapping rule learned as above. We then compare this binary representation $\mathbf{f}_y$ with the items in the database, or equivalently the rows of $\mathcal{C}$ using the minimum Hamming distance rule or any other relevant decoding method. As an example, consider the case when we have $|\mathcal{M}| = 8$ classes and in each class there are $p = 5$ instances with the data dimension $L = 2$ (chosen for the visualisation reasons). We consider a coding matrix $\mathcal{C}$ whose elements are based on *i.i.d.* realisations of a Bernoulli random variable $\mathcal{B}(p = \frac{1}{2})$. The matrix has 8 rows equal to the number of classes and we assign the number of columns to be $l = 4$. Therefore, the rate of identification defined as:

$$0 \leq R_{id} = \frac{\log_2 |\mathcal{M}|}{l} \leq 1 \tag{14}$$

which would be equal to $R_{id} = \frac{3}{4}$.

As for the training phase, we want to learn 8 independent SVM classifiers with the columns of $\mathcal{C}$ being the target values of training and the $8 \times 5 = 40$ total instances as the training examples. For each of the 4 classifications, the SVM rule considers almost half of the instances being targeted as belonging to the first class and the rest as belonging to the second class.

Figure 3 shows the class instances and Figure 4 shows the corresponding decision boundaries. As is clear from Figure 4, the decision boundaries follow the outer Voronoi cells of each class. This observation will serve us as a basis for the design of our proposed coding strategy in subsection 4.2.1.

4.2 Coding Matrix Design

An important concern in this method is the choice of the coding matrix. Using the techniques from coding theory the main focus of the researchers has been to design elements of $\mathcal{C}$ optimally to be able to combat against noise while keeping the rate of communication as close to 1 as possible, which implies fewer number of binary classifications. Through this end, most design strategies were in essence aiming at maximising the Hamming distance between the rows of $\mathcal{C}$.[10]

Another approach in the design of this matrix is to assign the elements randomly as we did in the example above. It is shown that this random assignment approach could do as good as the previous coding-oriented designs.[11,12]

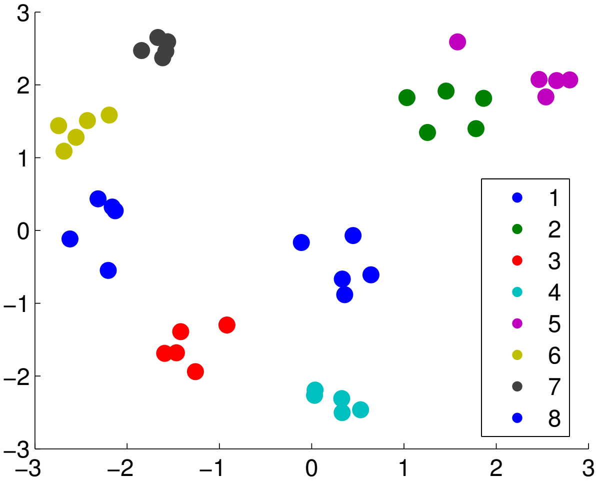


Figure 3. Instances of the above example with $|\mathcal{M}| = 8$ and $p = 5$.

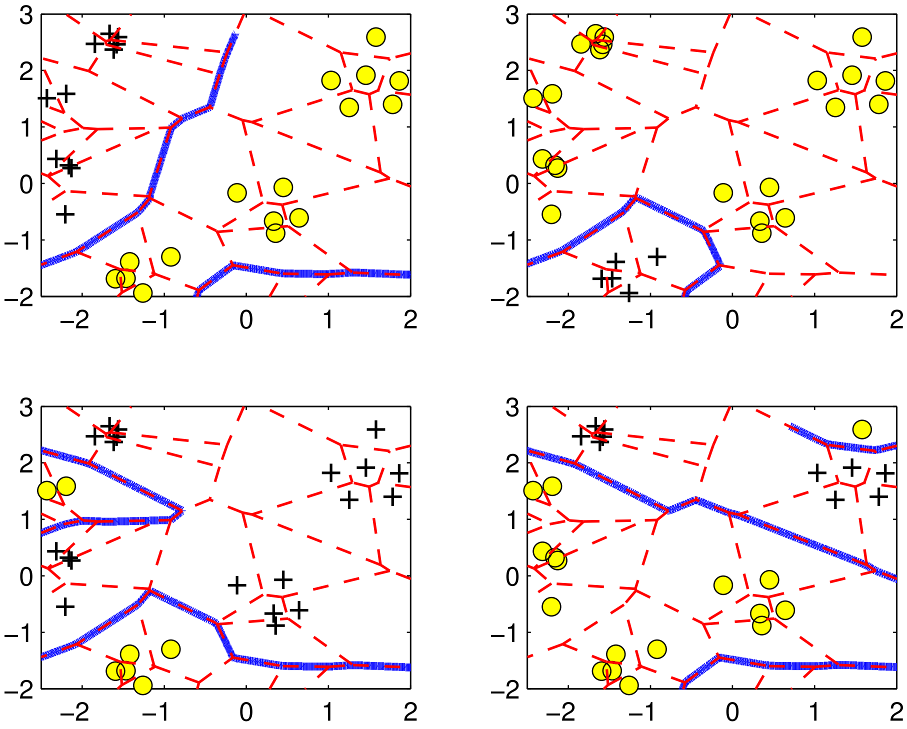


Figure 4. Equivalent Binary Classifications: Solid lines indicate the SVM decision boundaries and dashed lines indicate the Voronoi regions for each of the instances.

We considered this latter approach as one of the methods to design the coding matrix. Then in subsection 4.2.1, by formulating the problem differently, we will propose a novel optimal approach for coding matrix design.

It is important to mention that in the design of channel codes, e.g., famous codes like BCH or LDPC, it is generally assumed that the channel noise, in our case the bit flippings of $\mathbf{f}_y$ due to the classification process, is *i.i.d.*, or at least is independent of the input codewords. In the case of classification where we have a hypothetical channel representing the behaviour of the learning algorithm, however, we can observe that these bit flippings, or equivalently the channel noise samples are highly correlated with the input codewords[†]. Therefore, many of the information-theoretic explanations based on notions like typicality could also no longer be valid in this case. To address this issue, we consider this problem from a learning-theoretic viewpoint.

4.2.1 Proposed Coding Matrix Design

As it was shown in Figure 4, based on their large margin criterion, the SVM learned decision boundaries with nonlinear kernels are in fact approximations of the relevant Voronoi cells, i.e., the cells belonging to the instances of each of the two classes that are closest to each other. Notice that the choice of $\mathcal{C}$ dictates that in each of the l classifications, some of the adjacent classes be grouped together and labeled as '0' and some be grouped as '1' and also that this labelling and grouping could be changed in each classification.

Therefore, in order to design an optimal coding matrix, we ask the question: *Which choice of the elements of $\mathcal{C}$ guaranties learning these relevant Voronoi regions at least once, while avoiding redundant boundary learning*? The first requirement links to having good performance via optimal class separability and the latter implies maintaining the number of columns (classifiers) as small as possible, i.e., having maximal rate of communication.

Equivalently, the optimal assignment of codewords should work as a unique and non-redundant encoding of each class with the decision boundaries of SVM classifier following the Voronoi regions between the corresponding classes. To our knowledge such a design of coding matrix was not considered in the ECOC research.

This optimal design, in fact happens when the rows of $\mathcal{C}$ are chosen simply as the modulo-2 equivalent of the the class numbers, e.g., for a problem with $|\mathcal{M}| = 8$ classes, the optimal $\mathcal{C}$ is as shown in Figure 5(a).

$$\mathcal{C} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

(a) The Optimal Design

$$\mathcal{C} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

(b) A Random Design

Figure 5. Design of the optimal and a random coding matrix when $|\mathcal{M}| = 8$.

Intuitively, considering each of the two adjacent classes, this approach assigns each of them a codeword where they differ in at least one position. This satisfies our first objective because it guaranties that each pair of rows of $\mathcal{C}$ that could be assigned to two adjacent classes finds at least one position where the bit values are different. Therefore, they will be mapped into two different classes at least once which implies that the decision boundary will cross between them at least once.

With this approach, the second requirement is also fully satisfied. Since the length of each codeword, i.e, the number of columns of $\mathcal{C}$ equals $l = \log_2 |\mathcal{M}|$ which is minimal and the rate of communication is exactly 1.

As an example, consider the case when $|\mathcal{M}| = 16$ and to simplify the figure interpretation there is only $p = 1$ instance in each class with data dimension $L = 2$. Figure 6 shows the decision boundaries when we use the random coding technique described before with $l = 4$ classifiers.

[†] The formal proof of this property of noise is out of scope of this paper.

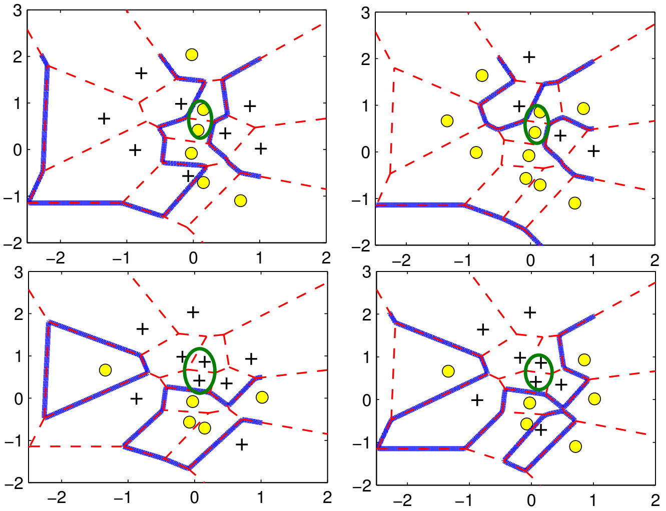


Figure 6. Decision boundaries with random coding ($|\mathcal{M}| = 16$, $p = 1$). No boundary separates two classes in the center that always leads to the classification error even in the absence of noise.

As is indicated by the ellipses in this figure, there is a pair of adjacent classes where in all the $l = 4$ classifications, their binary class labels happen to be the same. Therefore, the multiclass classification task is incomplete and we need to increase l to cope with this problem.

Figure 7 shows the same problem when it is treated with our proposed matrix design method. As is clear from the figure, all the adjacent pairs are targeted to different classes at least once.

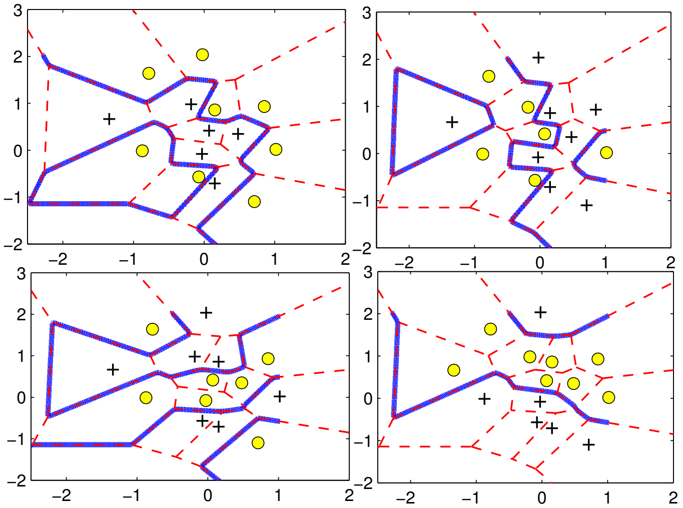


Figure 7. Decision boundaries with optimal coding matrix ($|\mathcal{M}| = 16$, $p = 1$).

In essence, the optimality of our approach is based on the fact that we learn the relevant Voronoi regions optimally. The fact that learning the Voronoi regions is the best that one could do is based on the maximum likelihood rule. It is important to mention that we are under the assumption that we are using SVM classifiers with nonlinear kernels that are properly tuned. Other learning rules do not necessarily learn these Voronoi regions and also an improper choice of kernel parameters does not guarantee learning these regions properly.

While ML decoding rule requires the knowledge of Voronoi regions for all classes, in our approach, these regions are distributed among the l binary classifiers and are learned implicitly. The fusion of the results of all binary classifiers, equivalently produces the entire coverage for all classes defined by their Voronoi regions. Therefore, the result of fusion of binary decodings should coincide with the optimal ML decoder.

The proposed approach is optimal in the ML decoding sense and should represent an interesting alternative to the existing binary classification rules based on more complex boundary learning strategies such as boosting.

5. PROPOSED APPROACH: BINARY CONTENT ENCODING

In this section, we go back to the identification problem. Being inspired by the materials in the previous section, here we consider the identification scenario essentially as a special case of the multiclass classification problem where in each class there is only one training example. Therefore, we have a total of $|\mathcal{M}|$ examples labeled as $\{1, 2, ..., |\mathcal{M}|\}$. This assumption, while making a natural link between the two areas, makes the analyses straightforward and also more practical when we have a huge number of classes.

In this method, as we did in the previous section for classification, we assign the rows of a binary coding matrix $C_{|\mathcal{M}| \times l}$ to each of the objects. Then we consider each of the columns of this matrix as the target values for binary classification whose inputs are the features in the real domain. We then train a separate classifier $g_i(\cdot) : \mathbb{R}^L \longmapsto \{1, 2\}$, for each of these l columns as shown in Figure 8.

Given a query item $\mathbf{y} \in \mathbb{R}^{\mathbb{N}}$, after the dimensionality reduction from N to L, we pass it through the l learned binary classifiers, $g_i(\cdot)$'s to give us a binary descriptor $\mathbf{b}_y$. Then we compare it with the rows of $\mathcal{C}$ to find the most similar object in the database which is assigned to that row. The similarity measure here could be the Hamming decoder or Bounded Distance Decoder. But if the query is not so distant from the objects in the database, or equivalently if the SNR is big enough, also depending on the choice of the coding matrix, the valid B_y's are always exactly equal to their respective row from the coding matrix. Hence, we will not need any decoding at all and the overall complexity of identification would be equal to the evaluation of the l learned functions.

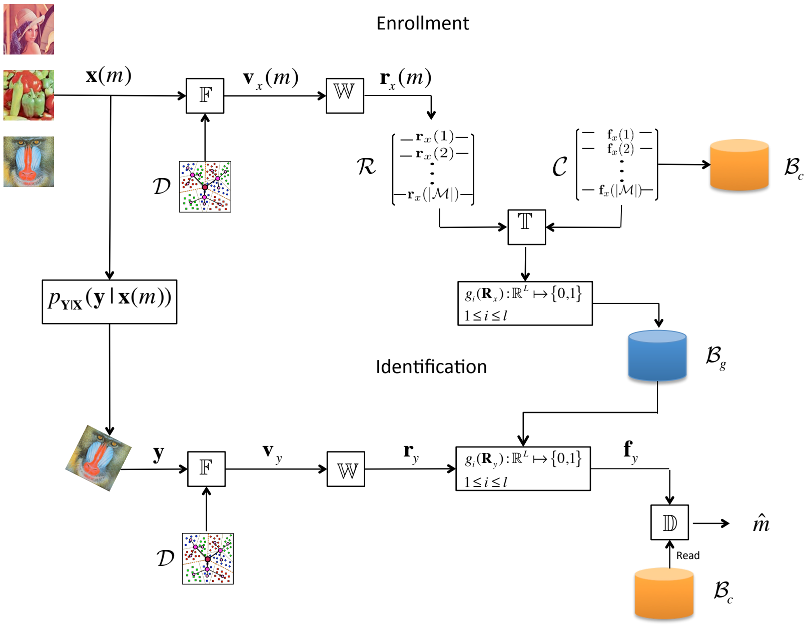

Enrollment
$\mathbf{x}(m)$
$\mathbb{F}$
$\mathbf{v}_x(m)$
$\mathbb{W}$
$\mathbf{r}_x(m)$
$\mathcal{D}$
$\mathcal{R}$
$\mathbf{r}_x(1)$, $\mathbf{r}_x(2)$, $\mathbf{r}_x(|\mathcal{M}|)$
$\mathcal{C}$
$\mathbf{f}_x(1)$, $\mathbf{f}_x(2)$, $\mathbf{f}_x(|\mathcal{M}|)$
$\mathcal{B}_c$
$\mathbb{T}$
$g_i(\mathbf{R}_x): \mathbb{R}^L \mapsto \{0,1\}$ $1 \leq i \leq l$
$\mathcal{B}_g$
$p_{\mathbf{Y|X}}(\mathbf{y} \mid \mathbf{x}(m))$
Identification
$\mathbf{y}$
$\mathbb{F}$
$\mathbf{v}_y$
$\mathbb{W}$
$\mathbf{r}_y$
$g_i(\mathbf{R}_y): \mathbb{R}^L \mapsto \{0,1\}$ $1 \leq i \leq l$
$\mathbf{f}_y$
$\mathbb{D}$
$\hat{m}$
$\mathcal{D}$
Read
$\mathcal{B}_c$

Figure 8. Identification based on Binary Content Encoding (proposed). $\mathbb{T}$ is the learning rule and $\mathbb{D}$ is the Hamming distance search.

5.1 Performance Analysis

As we have seen in sections 2.2 and 3.1, we could calculate the achievable rates of identification which equal the maximum possible mutual information between the inputs and outputs for those two cases. We then argued that the maximum number of distinguishable items in the database is bounded as $|\mathcal{M}| \leq 2^{LC_{GC}}$ and $|\mathcal{M}| \leq 2^{LC_{BSC}}$ for the real domain direct search and binary fingerprinting search, respectively. The capacity of the Gaussian Channel is an increasing function of SNR while the capacity of the BSC is bounded to 1 when SNR grows large. Figure 9 compares these capacities.

It should also be pointed out that the probability of correct identification could approach to one, only if $M \leq 2^{LC_{GC}}$ or $M \leq 2^{LC_{BSC}}$.

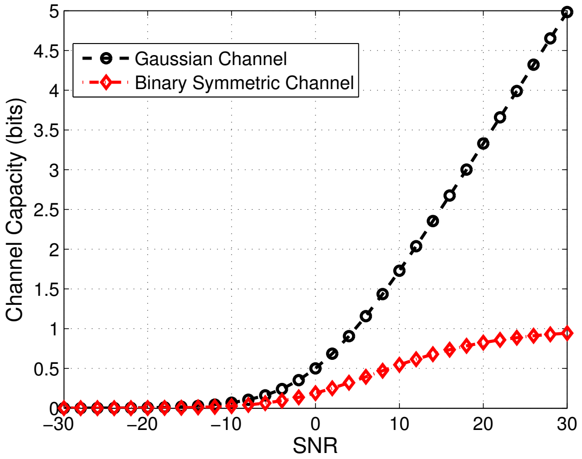


Figure 9. Comparison of identification capacities of Gaussian and binary symmetric channels.

The above analyses were based on the fact that the noise was considered to be independent and we could easily derive the channel capacities. In the learning approach, however, the situation is different and we cannot have a straightforward formula for capacity. But we can see that unlike in the fingerprinting case where $\mathbf{F}_y$ was only a function of $\mathbf{R}_y$, in the learning approach, it is also a function of $g_i(\cdot)$'s which are themselves functions of $\mathbf{R}_x$'s. Therefore, we expect that the capacity of identification for the proposed content encoding approach would be between the two curves in figure 9. By the optimal choice of the coding matrix as was proposed in subsection 4.2.1, we will see in the next section that the identification performance achieves its upper bound, that of the real domain approach.

6. EXPERIMENTAL RESULTS

We ran our experiments first on a synthetic dataset where we have more freedom to alter the parameters of the problem such as codebook size and SNR. Section 6.1 explaines these results. Then we validate our findings on the FAMOS dataset in section 6.2.

6.1 Synthetic Gaussian Data

We generate $\mathbf{x}(m) \in \mathbb{R}^N, \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbb{I}_N)$ and $1 \leq m \leq |\mathcal{M}|$ as our $|\mathcal{M}|$ enrolled objects in the database. The query items $\mathbf{y}$ is created as $\mathbf{Y} = \mathbf{X}(m) + \mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^N, \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbb{I}_N)$ is the additive noise. By fixing σ_X^2 and changing $SNR = 10 \log_{10} \frac{\sigma_X^2}{\sigma_Z^2}$ we control how the queries are distant from the original objects. In this part, because all the data is *i.i.d.*, we do not run any dimensionality reduction, therefore we have $L = N$.

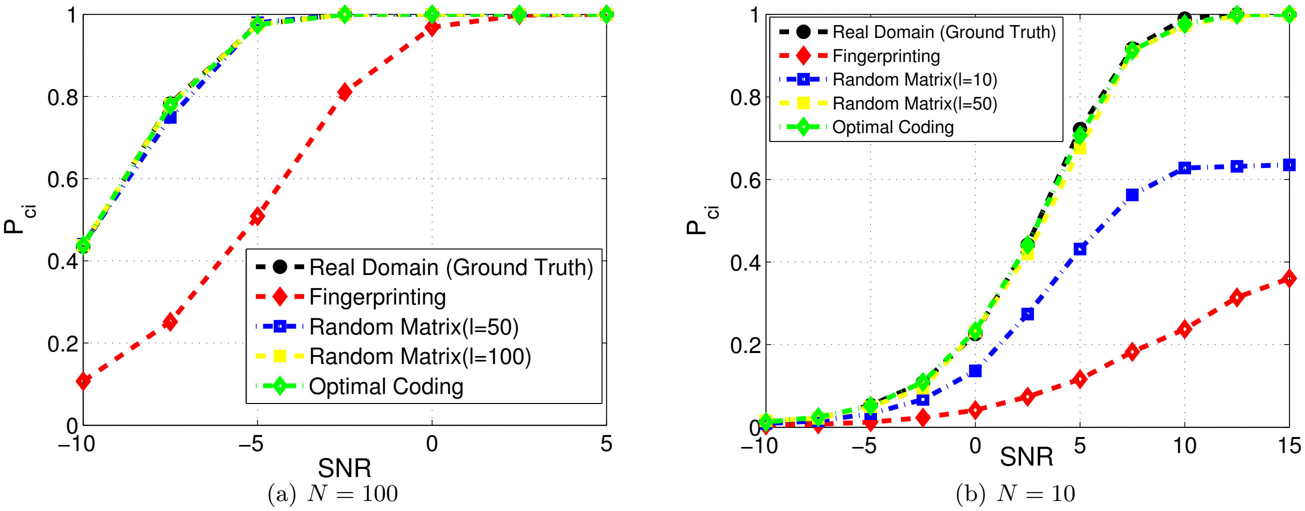
We experiment with the two explained coding matrix design approaches. For the random assignment case, the coding matrix $\mathcal{C}$ is a Bernoulli distributed, *i.i.d.* binary matrix with $\Pr\{\mathcal{C}(i,j) = 0\} = \Pr\{\mathcal{C}(i,j) = 1\} = \frac{1}{2}$

with $1 \leq i \leq |\mathcal{M}|$ and $1 \leq j \leq l$. For the optimal assignment, as was explained in 4.2.1, the matrix is chosen as the modulo-2 equivalents of m's from 0 to $|\mathcal{M}| - 1$.

We compare the performance of these methods with the fingerprinting approach as well as the direct comparison of the items in their original real domain using Euclidean distance search which is considered as the ground truth in terms of the probability of correct identification.

As for our learning algorithm, we use a simple implementation of the SVM classifier with Gaussian kernels.

Figure 10(a) illustrates the probability of correct identification with respect to changes in SNR when the codebook size and the data dimension were fixed to be $|\mathcal{M}| = 1024$ and $N = 100$ respectively. To see the behavior of random coding more clearly, we experimented with two different values for the number of classifiers which were $l = 50$ and $l = 100$. As we can see, by increasing the number of classifiers, the random coding approach improves in performance. However, this improvement comes with the cost of adding to the amount of training time while in our optimal coding approach we get the upper limit of performance with the least possible number of classifiers which in this case is $l = 10$.



(a) $N = 100$

(b) $N = 10$

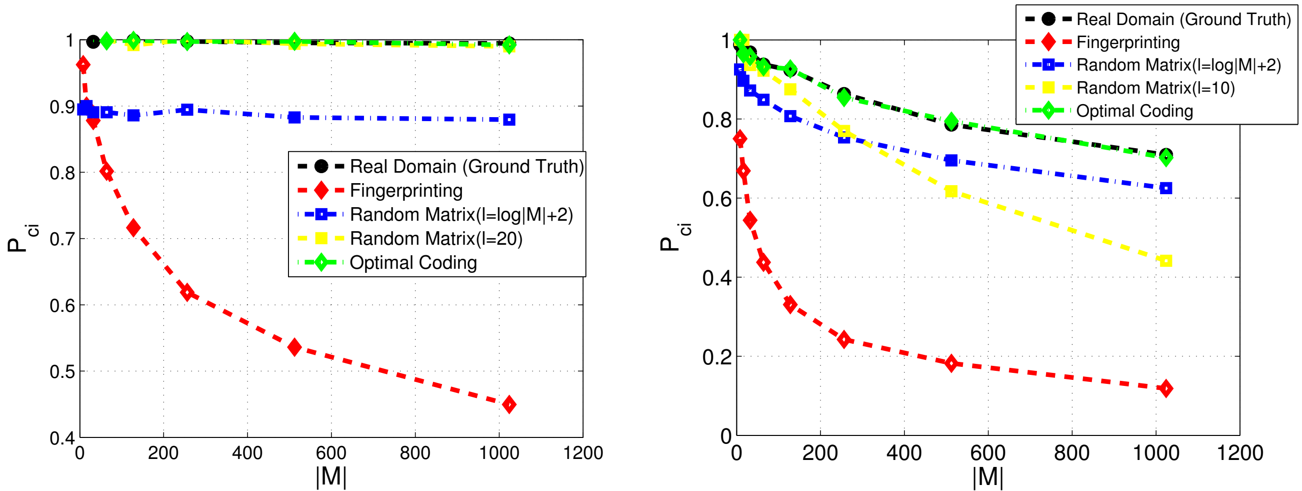
Figure 10. Identification performance of the synthetic Gaussian data with $|\mathcal{M}| = 1024$ and variable SNR.

In Figure 10(a) the data dimension $N = 100$ was high enough, so that for SNR's greater than -3.5 dB which results in $C_{BSC}(SNR = -3.5) = 0.1045$ bit, the maximum achievable database size was higher than our chosen $|\mathcal{M}|$, i.e., $|\mathcal{M}| = 1024 < 2^{L \times 0.1045} \simeq 1398$. Therefore, we expect to have good performance for SNR's greater than this threshold in all methods, even the fingerprinting approach. However, this situation is not always the case. Depending on the parameters of the problem, we might achieve the identification rate in the real domain while always below the capacity in the fingerprinting approach for all values of SNR. Figure 10(b) demonstrates one of these cases. The parameters of this experiment where chosen as above, except for the data dimension which was chosen as $N = 10$. In this case, for SNR's greater than 5 dB resulting in $C_{BSC} = 0.3579$ and $C_{GC} = 1.029$, the maximum database size in the real domain is calculated as $2^{L \times 1.029} \simeq 1252$ which is bigger than our experimented size, while for the fingerprinting case, we will not be able to identify $|\mathcal{M}| = 1024$ items reliably for no value of SNR. Therefore, as is clear from the figure, the performance of the fingerprinting approach is always restricted. Also the random coding approach with $l = 10$ which in this case both equals to the data dimension and $\log_2 |\mathcal{M}|$ fails to achieve a good performance because it needs more classifiers to be trained while the optimal approach, with the same amount of classifiers achieve a perfect performance.

Now we experiment with variable database sizes. Figure 11(a) illustrates the probability of correct identification versus the database size when we have $SNR = 5$dB fixed for all experiments. The data dimension is

$N = 20$. In order to make a comparison with both the fingerprinting approach and the optimal assignment, we choose the number of classifiers in the random coding method once equal to the data dimension, i.e, $l = L$ and also once near its minimum value respectively.

Figure 11(b) presents the results of the same experiment when we change the data dimension to $N = 10$. As we see in this case, as well as all other cases, the optimal design achieves the performance as good as the ground truth.



(a) $N = 20$

(b) $N = 10$

Figure 11. Identification performance of the synthetic Gaussian data with fixed $SNR = 5$ and variable $|\mathcal{M}|$

6.2 FAMOS Dataset

To compare the algorithms on real data, we tested the algorithms on the FAMOS dataset which contains 5000 unique microstructures from consumer packages and is used for the development, testing and benchmarking of forensic identification and authentication technologies.

We tested the probability of correct identification for all of the four algorithms, again with two different l's for the random assignment approach, versus the size of the reduced dimension after random projection based dimensionality reduction. Because we are not using any feature extraction method and we consider the image pixel values directly as our data elements which are vectorised versions of the 128×128 images, these values tend to be highly redundant and they also have a big dimension. Therefore, dimensionality reduction is a necessary step.

Figure 12 shows the results of identification when using Camera1-A in the database as our original objects and Camera2-B as our query items. As we see from the figure, for dimensions reduced to more than about 70, which is much smaller than the original dimensionality, we achieve perfect identification. This implies that the dimensionality reduction based on random projections was indeed successful. Also the performance of our proposed optimal coding approach applied to a real dataset, again achieves the upper bound. We can also see that the random coding approach only achieves good performance when it is trained with a big number of classifiers.

7. SUMMARY

In this work we investigated different approaches for the problem of content identification. To this end, we proposed a machine learning oriented framework to address the problem by considering the identification scenario as multiclass classification. We advocated a coding theoretic approach to solve this later problem which aims at the efficient use of binary classifiers to solve the multiclass case. We then proposed a novel optimal method

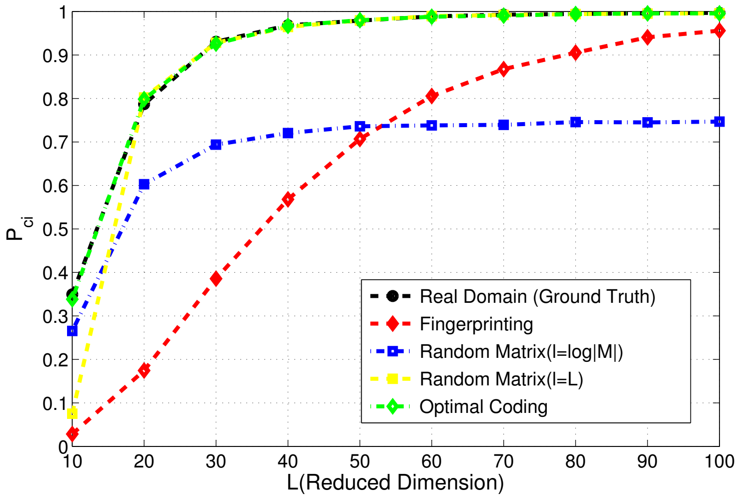


Figure 12. The performance of identification methods on the FAMOS dataset with $|\mathcal{M}| = 5000$ items.

of coding matrix design to be used in this framework. To analyse the problem in detail and gain more insight, we tested the performance of these methods first on a synthetic dataset. We also tested the methods on a real database of physical microstructures. In all the experiments, we could see that our proposed machine learning based approach, especially when treated with the proposed coding matrix design, achieves the upper bounds of performance in terms of the probability of correct identification while using the minimum possible number of binary classifiers. To be investigated in a future work, we also argue that, due to the optimal use of prior data in machine learning methods, our approach has this important advantage to incorporate the data efficiently from more than one observation of the same object in the database. This could significantly boost the identification performance and robustness.

ACKNOWLEDGMENTS

This work has been partially supported by the research project PSPB-125/2010. The authors are thankful to Mr. Maurits Diephuis for his comments on the paper and discussion about the machine learning approach to the content identification.

REFERENCES

[1] Voloshynovskiy, S., Koval, O., Beekhof, F., and Pun, T., "Conception and limits of robust perceptual hashing: toward side information assisted hash functions," in [*Proceedings of SPIE Photonics West, Electronic Imaging / Media Forensics and Security XI*], (2009).

[2] Voloshynovskiy, S., Koval, O., Beekhof, F., and Pun, T., "Robust perceptual hashing as classification problem: decision-theoretic and practical considerations," in [*Proceedings of the IEEE 2007 International Workshop on Multimedia Signal Processing*], (October 1–3 2007).

[3] Varna, A. and Wu, M., "Modeling and analysis of correlated binary fingerprints for content identification," *Information Forensics and Security, IEEE Transactions on* **6**(3), 1146–1159 (2011).

[4] 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).

[5] Farhadzadeh, F., Voloshynovskiy, S., and Koval, O. J., "Performance analysis of content-based identification using constrained list-based decoding," *IEEE Transactions on Information Forensics and Security* **7**(5), 1652–1667 (2012).

[6] Farhadzadeh, F., Voloshynovskiy, S., Koval, O., and Beekhof, F., "Information-theoretic analysis of content based identification for correlated data," in [*IEEE Information Theory Workshop (ITW)*], 205–209 (2011).

[7] Cover, T. M. and Thomas, J. A., [*Elements of Information Theory 2nd Edition*], Wiley-Interscience, 2 ed. (7 2006).

[8] Voloshynovskiy, S., Koval, O., Beekhof, F., Farhadzadeh, F., and Holotyak, T., "Information-theoretical analysis of private content identification," in [*IEEE Information Theory Workshop, ITW2010*], (Aug.30-Sep.3 2010).

[9] Dietterich, T. G. and Bakiri, G., "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res. (JAIR)* **2**, 263–286 (1995).

[10] Murphy, K. P., [*Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*], The MIT Press (8 2012).

[11] James, G. and Hastie, T., "The error coding method and picts," *J. Computational and Graphical Statistics* **7:3**, 377–387 (1998).

[12] Voloshynovskiy, S., Koval, O., Beekhof, F., and Holotyak, T., "Information-theoretic multiclass classification based on binary classifiers," *Signal Processing Systems* **65**, 413–430 (2011).