Content-Identification : towards Capacity Based on Bounded Distance Decoder

Farhadzadeh, Farzad; Voloshynovskyy, Svyatoslav; Koval, Oleksiy

Content-Identification: towards Capacity Based on Bounded Distance Decoder

Farzad Farhadzadeh, Sviatoslav Voloshynovskiy, and Oleksiy Koval

University of Geneva, Stochastic Information Processing (SIP) Group

Abstract

In recent years, content identification based on digital fingerprinting attracts a lot of attention in different emerging applications. In this paper, we perform the information–theoretic analysis of finite length digital fingerprinting systems. We show that the identification capacity over the Binary Symmetric Channel (BSC) is achievable by using Bounded Distance Decoder (BDD).

1 Introduction

In the last 10 years, content identification based on digital fingerprinting performed an impressive evolution from just an alternative solution to digital watermarking in copyright protection to a standalone domain of research. A digital fingerprint represents a short, robust and distinctive content description allowing fast and privacy-preserving operations. In this case, all operations are performed on the fingerprint instead of on the original large and privacy–sensitive data.

Some important practical and theoretical achievements were reported during last years. The main efforts on the side of practical algorithms have been concentrated on the robust feature selection and fast indexing techniques mostly borrowed from content-based retrieval applications [1], [2]. The information–theoretic limits of content identification under infinite length and ergodic assumptions have been investigated by Willems et. al. [3] using the jointly typical decoder.

In this paper we introduce the information–theoretic framework for the analysis of content identification system based on finite length fingerprinting. Contrary to previous works, we consider alternative decoding rules and demonstrate their capability to achieve the identification capacity limit under asymptotic assumptions.

2 Identification Setup

The identification problem is to determine whether a query $\mathbf{y}$ is related to some element of the database, and if so, to identify which one. To this end an algorithm $\psi(\mathbf{y})$ must be designed, returning:

$$\psi(\mathbf{y}) \in \{0, 1, 2, \ldots, 2^{NR}\} \tag{1}$$

where N is the database entry length, R is an identification rate and the decision $\psi(\mathbf{y}) = 0$ indicates that $\mathbf{y}$ is unrelated to any database elements. In this paper, we consider the identification problem for binary fingerprinting schemes. This scenario can be modeled as a multiple hypothesis test:

$\mathcal{H}_0$**:** The query $\mathbf{Y}$ is unrelated to any database entry,

$\mathcal{H}_m$**:** The query $\mathbf{Y}$ is related to the m^{th} entry of the database.

We assume that the fingerprint bits are i.i.d. and equally likely to be 0 and 1, and the probabilistic mismatch between the channel input $\mathbf{X}$ and output $\mathbf{Y}$ is modeled by the BSC.

The decoder performance can be evaluated in terms of:

- Probability of false positive ($P_{FP} \triangleq \Pr\{\psi(\mathbf{y}) > 0 | \mathcal{H}_0\}$),
- Probability of false negative($P_{FN} \triangleq \sum_{m=1}^{2^{NR}} \Pr\{\psi(\mathbf{y}) \neq m | \mathcal{H}_m\}$).

2.1 The Decoding Rule

Under this model, upon receiving a query $\mathbf{y}$, the Bounded Distance decision rule can be defined as:

$$\psi(\mathbf{y}) = \begin{cases} \tilde{m}, & \text{if for a unique } \tilde{m},\ d_H(\mathbf{x}(\tilde{m}), \mathbf{y}) \leq \gamma N, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where γ is the normalized threshold and $d_H(\cdot, \cdot)$ denotes the Hamming distance.

3 Error Exponents

In this section, we derive bounds on the probabilities of false positive and negative using i.i.d. binary fingerprints with the decision rule given by (2).

The probability of false positive is given by:

$$\begin{aligned} P_{FP} &= \Pr\left\{\bigcup_{m=1}^{2^{NR}} d_H(\mathbf{Y},\mathbf{X}(m)) \le \gamma N|\mathcal{H}_0\right\} \overset{(a)}{\le} \sum_{m=1}^{2^{NR}} \Pr\{d_H(\mathbf{Y},\mathbf{X}(m)) \le \gamma N|\mathcal{H}_0\} \\ &\overset{(b)}{=} 2^{NR}\Pr\{d_H(\mathbf{Y},\mathbf{X}(1)) \le \gamma N|\mathcal{H}_0\}, \end{aligned} \tag{3}$$

where (a) follows from the union bound, and (b) follows from the fact that the fingerprints $\mathbf{X}(m)$ are i.i.d. . Under $\mathcal{H}_0$, $\mathbf{Y}$ and $\mathbf{X}(1)$ are independent, thus $d_H(\mathbf{Y},\mathbf{X}(m)) \sim \mathcal{B}(N,\frac{1}{2})$. By using the Chernoff–bound on the tail of binomial distributions $\mathcal{B}(N,\frac{1}{2})$, the probability of false positive can be bounded as:

$$P_{FP} \le 2^{NR}2^{-Nd(\gamma\|\frac{1}{2})} = 2^{-N(1-H_2(\gamma)-R)}, \tag{4}$$

where $d(\alpha\|\beta) \triangleq \alpha\log_2\frac{\alpha}{\beta} + (1-\alpha)\log_2\frac{1-\alpha}{1-\beta}$ is defined as the divergence between $0 \le \alpha \le 1$ and β and $H_2(\gamma)$ denotes the binary entropy.

Now consider the probability of false negative P_{FN}. Given that $\mathcal{H}_m, m \neq 0$ is true, the false negative event occurs if $d_H(\mathbf{Y},\mathbf{X}(m)) > \gamma N$ or if there is any $m' \neq m$ such that $d_H(\mathbf{Y},\mathbf{X}(m')) \le \gamma N$. Thus the probability of false negative is defined by:

$$\begin{aligned} P_{FN} &= \Pr\left\{d_H(\mathbf{Y},\mathbf{X}(m)) > \gamma N \cup \bigcup_{m'\neq m} d_H(\mathbf{Y},\mathbf{X}(m')) \le \gamma N|\mathcal{H}_m\right\} \\ &\overset{(a)}{\le} \Pr\{d_H(\mathbf{Y},\mathbf{X}(m)) > \gamma N|\mathcal{H}_m\} + \sum_{m'\neq m}\Pr\{d_H(\mathbf{Y},\mathbf{X}(m')) \le \gamma N|\mathcal{H}_m\} \\ &\overset{(b)}{\le} 2^{-Nd(\gamma\|P_b)} + 2^{-N(1-H_2(\gamma)-R)}, \end{aligned} \tag{5}$$

where (a) follows from the union bound, and (b) follows from Chernoff–bound and the facts that under $\mathcal{H}_m$, $d_H(\mathbf{Y},\mathbf{X}(m)) \sim \mathcal{B}(N,P_b)$, where P_b is the crossover probability of the BSC, $d_H(\mathbf{Y},\mathbf{X}(m')) \sim \mathcal{B}(N,\frac{1}{2})$.

In order to consider the probabilities of false positive and negative togethers, the summation of P_{FP} and P_{FN} is bounded. By combining (4) and (5), this bound is given by:

$$P_{FP} + P_{FN} \le 2^{-Nd(\gamma\|P_b)} + 2^{-N(1-R-H_2(\gamma)-\frac{1}{N})}. \tag{6}$$

One can conclude that for $Pb \le \gamma \le \frac{1}{2}$ and if $H_2(\gamma) \le 1-R-\frac{1}{N}$, there exist fingerprints with rate R such that $\lim_{N\to\infty}(P_{FP}+P_{FN}) = 0$. Since

this holds for γ arbitrarily close to P_b, we claim that $R = 1 - H_2(P_b)$ is achievable and from [3], this is the identification capacity setup.

We now investigate the "best" upper bound over the probability of false. It is explicit that

$$P_{FP} + P_{FN} \leq 2 \times 2^{-N \min[d(\gamma \| P_b), (1 - H_2(\gamma) - R - \frac{1}{N}))]}. \tag{7}$$

Therefore, the best bound is achieved by γ_{opt}, which makes both terms equal. The optimum value of threshold is given by:

$$\gamma_{opt} = \frac{1 - R + \log_2(1 - P_b) - \frac{1}{N}}{\log_2 \frac{1-P_b}{P_b}}. \tag{8}$$

Note that as $N \to \infty$, γ_{opt} converges to the optimum threshold under communication setup [4].

4 Conclusion

In this paper, we have presented a framework for content identification based on binary fingerprints. By performing an information-theoretic analysis of finite length digital fingerprinting systems and proved that by using the BDD, the identification capacity limit under asymptotic assumptions is achievable.

Acknowledgement

This paper was partially supported by SNF project 200021-119770.

References

[1] J. Haitsma, T. Kalker, and J. Oostveen,"Robust audio hashing for content identification," in International Workshop on CBMI 2001.

[2] F. Lefebvre and B. Macq, Rash : RAdon Soft Hash algorithm, in Proc. of EUSIPCO, Toulouse, France, 2002.

[3] F. Willems, T. Kalker, J. Goseling, and J. Linnartz, "On the capacity of a biometrical system", in Proc. 2003 IEEE ISIT, Yokohama, Japan.

[4] https://www.sps.ele.tue.nl/members/F.M.J.Willems.