Collusion Secure Convolutional Fingerprinting Information Codes

Yan Zhu
Institute of Computer Science and Technology
Peking University, 100871
Beijing, China
zhuyan@icst.pku.edu.cn

Wei Zou
Institute of Computer Science and Technology
Peking University, 100871
Beijing, China
zouwei@icst.pku.edu.cn

Xinshan Zhu
Institute of Computer Science and Technology
Peking University, 100871
Beijing, China
zhuxinshan@icst.pku.edu.cn

ABSTRACT

Digital Fingerprinting is a technique for the merchant who can embed unique buyer identity marks into digital media copy, and also makes it possible to identify "traitors" who redistribute their illegal copies. At present, the fingerprinting scheme generally have many difficulties and disadvantages for large-size uses problems involve in the code construction with shorter length and effective traitor tracing. To resolve these problems, this paper presents the definition of Fingerprinting Information Code and a practical construction method by composing of convolutional codes and generally fingerprinting codes based on Boneh-Shaw model. Its decoding algorithm is presented by introducing the ideal of 'Optional Code Subset' and improving Viterbi algorithm. The security properties and performance are proved and analyzed by theory and example. As the results, the proposed scheme has shorter information encoding length and achieves optimal traitor searching in larger number of buyers.

Key words:

Categories and Subject Descriptors

H.1.1 [**Information Systems**]: Models And Principles—*Systems and Information Theory*

General Terms

Security, Performance

Keywords

digital fingerprinting, collusion security, tracing traitor, convolutional code

1. INTRODUCTION

Digital fingerprinting is a technique for the merchant who can embed unique buyer identity marks into digital media copy, and also makes it possible to identify 'traitors' who redistribute their illegal copies by obtaining the sellers' information from the redistributed contents. Such unique marks of the buyer are called fingerprint. Digital fingerprinting, is initially considered only as a coding technique for resisting collusion attacks and tracing traitor, is implementing some new requirements to enhance the copyright protection power includes such properties as imperceptive, undeniable, anonymous and so on. At present, digital fingerprinting has been a synthesis technology that involves signal process, code theory and cryptography. In the traditional fingerprint construction, a merchant chooses randomly a codeword for each buyer, and then he can identify traitor by this codeword after collusion attacks. However, there exists some inconveniences for high-level application because such method does not achieve the encoding of information. To implement this function, the aim of this paper is to construct fingerprint information codes with collusion-secure and tracing traitor by hiding the information of the buyer into digital multimedia.

Boneh and Shaw first present a relatively rounded concept and an explicit construction of fingerprint codes (called BS model) for collusion secure [1]. They have shown that apart from some trivial cases the codes cannot be constructed which guarantee absolute security for innocent buyers. But allowing that a innocent person comes under suspicion with probability ε they constructed c-secure codes with ε-error which demand polynomially in $log(1/\varepsilon)$ and $log(n)$ many different marking positions, where n is the number of possible buyers. However, the tracing algorithm in model is NP-hard problem. To resolve this problem, Barg and Blakly *et al* construct a code with Identifiable Parent Property (IPP) based on algebraic codes to reduce decoding complexity to $ploy(n)$ [2]. However, the codes is only suitable for the case of size 2 coalitions, either one of the traitors is identified with probability 1 or both traitors are identified with probability $1-\exp(\Omega(n))$. Furthermore, with respect to multimedia such as video and audio, some scholars have presented the random fingerprint coding methods. For example, Wang proposed a fingerprinting algorithm and the corresponding tracing algorithm by using a pseudo-random sequence to control the embedding of the fingerprint bits [3]. But this algorithm requires to save every secret-key for the pseudo-random number generator so that it is unpractical to tracing traitors with large user number.

ASIACCS'06, March 21-24, 2006, Taipei, Taiwan.

It is very important to provide a rapid, accurate and cost effective fingerprinting scheme for identifying the traitors. To achieve such a construction, above all, this paper gives a precise definition of the fingerprinting problem with the aim to establish relations between different levels of the problem: tracing ability and information coding. Secondly, we propose a construction of fingerprinting information codes. The construction employs the concatenation of two codes: the inner fingerprint code and the outer convolutional code. Furthermore, considering the usual fingerprint codes have the capacity to trace many codewords of coalition, Viterbi algorithm of convolutional code is improved by Optional Code Sets. Furthermore, an identification algorithms of complexity polynomial in the length of the codes are proposed, and then the properties of code length, collusion security are proved and analyzed by the theory of error-correcting codes. As the results, this coding scheme is easy to implement and has lower complexity. Moreover, it achieves a shorter fingerprinting length and more optimal traitor searching.

The remainder of this paper is organized as follows. Section 2 introduces the concept of fingerprinting information codes. Section 3 gives a brief description on Boneh and Shaw model. Section 4 describes our convolutional fingerprinting information codes as well as its encoding and decoding algorithms. The Performances are proved and analyzed in section 5. Finally, section 6 gives conclusion.

2. DIGITAL FINGERPRINTING INFORMATION CODES

Without loss of generality, digital fingerprinting scheme generally involves four phases: information coding, mark embedding, mark detection and tracing traitor. Here, the embedding of one symbol is called a mark among the fingerprint. The fingerprinting codes is composed of an algorithm pair (E, T) and a symbol sequence set $\Gamma \subseteq \Sigma^l$ in a symbol set Σ with s symbols, where E is a generation algorithm to generate fingerprint for user u with correlation information r, that is $I = E(u, r) \in \Gamma$, and T is a tracing algorithm to recognise the illegal forger with codeword $I' \in \Sigma^l$ and distribution register, that is $u = T(I', m)$, where u also can be a forger set.

In order to embed the fingerprint of user u_i to a original object O, at first, we apply E to generate a fingerprint $I_i = (i_1, i_2, \cdots, i_L)$, i.e. $I_i = E(u_i)$, where $i_j \in \Sigma$ $(1 \le j \le L)$. Secondly, the object O is divided into many blocks and we randomly choose L blocks into sequence $O = (o_1, o_2, \cdots, o_L)$. Next, embedding algorithm creates a copy $O^{(u_i)}$ by embedding i_j into o_j. When a pirated copy is found, the detection algorithm extracts the marks from every block and outputs all the fingerprint $I' = (i'_1, i'_1, \cdots, i'_L)$. Finally, the tracing algorithm T identifies the coalition $C = \{u_1, u_2, \cdots, u_c\}$ according to the record M, i.e. $C = T(I', M)$.

It is inevitable for fingerprinting to make the same content contain different marks. Let $C = \{u_1, u_2, \cdots, u_c\}$ is a coalition of c users who hold many copies with fingerprint. It is a feasible attack policy for a coalition of users to detect specific marks if they differs between their copies, and then the colluders construct a illegal copy by selecting different blocks of marks from among their content and piecing the new blocks together. Another kind of cost-effective attack is a process in which several differently marked copies of the same content are averaged to disrupt the underlying watermarks. These attacks by a coalition of users with the same content containing different marks are called 'Collusion Attack'. A fingerprinting scheme is required to assign unique codewords for copyright protection on the basis of collusion resistance, which is called collusion security fingerprinting code. Such a code is generally defined as follows:

Definition 1. (Fingerprinting code) (l, n)-fingerprinting scheme is a function $E(u)$ which maps a user serial number u $(1 \le u \le n)$ to a codeword in Σ^L, where Σ is an alphabet. When a coalition of at most c users, $C = \{u^1, u^2, \cdots, u^c\}$, employs $(E(u^1), E(u^2), \cdots, E(u^c))$ to create a word $z \in \Sigma^L$, this code is called c-secure with ε-error if there exists a tracing algorithm A that finds at least a user with probability of at least $1 - \varepsilon$, that is $Pr\{A(z) \in C\} \ge 1 - \varepsilon$, where probability is taken over the random choice of A and coalition C.

Besides collusion resistance, most of the applications should expect to encode the user's information into fingerprint. This kind of code, called the fingerprinting information code, is usually constituted by concatenation code on the basis of the fingerprinting code [4]. A fingerprinting information code is generally defined as follows:

Definition 2. (Fingerprinting information code) (L, N)-fingerprinting scheme is a function $\Phi(m^u, r)$ which maps a user information m^u $(1 \le u \le N)$ to a codeword in Σ^L, where Σ is an alphabet. When a coalition of at most c users, $C = \{u^1, u^2, \cdots, u^c\}$, employs $C' = (\Phi(m^{u^1}, r^1), \Phi(m^{u^2}, r^2), \cdots, \Phi(m^{u^c}, r^c))$ to create a word $z \in \Sigma^L$, this code is called c-secure with ε-error if there exists a tracing algorithm A that finds at least a code $\Phi(m^{u^i}, r^i)$ for C' with probability of at least $1 - \varepsilon$, that is $Pr\{A(z) \in C'\} \ge 1 - \varepsilon$, where probability is taken over the random choice of Φ, coalition C and string $r^i (1 \le i \le c)$.

3. BONEH-SHAW FINGERPRINTING MODEL

In [1], Boneh-Shaw presents a construction of c-frameproof code by the composition of c-frameproof (l, p)-code Γ and $(L, N, D)_p - ECC$ code Θ under the collusion attack assumption. At first, in order to model the strategy of forge codewords in any coalition of c users, 'marking assumption' is defined and then it is requires that the fingerprint codes could endure any attack under such assumption. Let $\sum$ denote an finite alphabet of size s representing the s different states of the marks, in which each symbol will be denoted by the integers 1 to s. A set $\Gamma = \{w^{(1)}, w^{(2)}, \cdots, w^{(n)}\} \subseteq \sum^l$ will be called an (l, n)-code, where every codeword $w^{(i)} = \{w_1^{(i)}, w_2^{(i)}, \cdots, w_l^{(i)}\}$ will be assigned to user u_i, for $1 \le i \le n$. We refer to the set of words in Γ as the codebook. Marking assumption can be defined as follow:

Definition 3. (Marking Assumption) Let $\Gamma = \{W^{(1)}, W^{(2)}, \cdots, W^{(n)}\}$ is an (l, n)-code and $C = \{u_1, u_2, \cdots, u_c\}$ is a

coalition of c-traitors. Let we say that position i is undetectable for C if the words assigned to users in C match in i'th position, that is $w_i^{(u_1)} = \cdots = w_i^{(u_c)}$. For detectable position, we define the feasible set Γ of C as

$$\Gamma(C) = \{x = (x_1, \cdots, x_l) \in \Sigma^l | x_j \in w_j, 1 \le j \le l\} \quad (1)$$

where

$$w_j = \begin{cases} \{w_j^{(u_1)}\} & w_j^{(u_1)} = \cdots = w_j^{(u_c)} \\ \{w_j^{(u_i)} | 1 \le i \le c\} \cup \{\perp\} & otherwise \end{cases} \quad (2)$$

where $\perp$ denotes an erased mark.

An (l, n) fingerprinting scheme is a function $\Gamma(u, r)$ which maps a user number $1 \le u \le n$ and a string of random bits $r \in \{0, 1\}^*$ to a codeword in Σ^l. The random string r is the set of random bits used by the distributor and keeps hidden from the users. The model denotes a fingerprinting scheme by Γ_r. The security of such scheme is defined as follows:

theorem 1. For $n \ge 3$ and $\varepsilon > 0$, let $d = 2n^2 log(2n/\varepsilon)$. The fingerprinting scheme $\Gamma_0(n, d)$ is n-secure with ε-error.

The length of Γ_0 is linear in the number of users and it is therefore impractical. Hence, BS model uses the code Γ_0 to construct shorter codes. A concatenation code is defined as follows:

Definition 4. A set $\Re$ of N words of length L over an alphabet of p letters is said to be an $(L, N, D)_p$-Error Correcting Code or in short, an $(L, N, D)_p - ECC$, if the Hamming distance between every pair of words in $\Re$ is at least D.

In BS model, Let Γ' is the composition of (l, n)-code Γ and $(L, N, D)_p - ECC$ code $\Re$. Then the code Γ' is an (lL, N)-code and it is a c-frameproof code. The codeword in Γ' is uniform random distribution. The following theorem provides the security property of Γ'.

theorem 2. Given integers N,c and $\varepsilon > 0$, set $n = 2c$, $L = 2c \log(2N/\varepsilon)$ and $d = 2n^2 \log(4nL/\varepsilon)$. Then, $\Gamma'(L, N, n, d)$ is a code which is c-secure with ε-error. The code contains N words and has length $l = O(Ldn) = O(c^4 \log(N/\varepsilon) \log(1/\varepsilon))$.

Where, n is the number of codeword and each bit is duplicated d times in each block for Γ code.

Although BS model resolves the construction problem of shorter codes from the theory of error correcting codes, the performance of tracing algorithm is worse since it finds a member of the guilty coalition by traversing one by one and merely arbitrarily chooses one of the outputs of inner algorithm for each component [5]. In order to resolve these problems, we focus on the following two aspects:

- Considering its collusion-resistant mechanisms, the fingerprint code has sufficient capability to find more than one member of the coalition. We should take full advantage of the capability.
- Since the tracing problem is known to be a NP-hard problem, we should propose a fingerprinting scheme that achieve a effective decoding algorithms with polynomial-time complexity relying on coding theory [6].

4. CONVOLUTIONAL FINGERPRINTING SCHEME

The target of this paper is that constructs a ε-error and c-secure fingerprinting scheme for N-users to improve runtime and reduce storage. For this purpose, this section presents a convolutional fingerprinting scheme by using the capacity of tracing more than one codeword in the coalition. Therefore, we replace the general Error Correcting Codes in BS model with convolutional codes. The scheme employs two-layer structure by composing an inner frameproof code with an outer convolutional code. The major differences between the presented scheme and classical BS model are:

1. The inner codes are constructed by frameproof codes in order to achieve collusion-resistent, multiple value output.
2. The outer codes are constructed by convolutional codes to reduce codeword length, and the decoding efficiency is improved by Maximum Likelihood Decoding algorithm. It is realized that the fingerprinting information length is irrelative to user size and inner codeword length.

The scheme is illustrated from the encoding and decoding process as follows.

4.1 Convolutional fingerprinting encoding

Convolutional error correcting codes were first introduced by Elias and are widely applied today in telecommunication systems, e.g., radio, satellite links, mobile communication [8]. Convolutional codes differ from the block codes in that each encoding operation depends on current and a number of previous information groups. Convolutional code has some advantage to build self-orthogonal code and punctured code. Therefore, convolutional decoding can be performed using a Viterbi algorithm which is the more convenient to obtain the optimum decoding than block codes. Hence, the construction of fingerprinting scheme is based entirely on the convolutional code.

The presented $\Phi(L, N, l, n)$-fingerprint code has two-layer concatenate structure: the outer layer is $\Im(n_0, k_0, m_0)$-convolutional code called Convolutional Error-Correcting Layer and the inner layer is $\Gamma(l, n)$ code called Fingerprint Layer, where N is the user number and L is convolutional code length. A convolutional code group is called as a fingerprint word. Let c is the maximum collusion number, $m^{(u_i)}$ denotes the identification symbol string of user u_i $(1 \le i \le N)$. The $m^{(u_i)}$ is assigned randomly with uniform distribution and assures unique to each user. Let an (n_0, k_0, m_0) convolutional encoder over the Galois field $GF(2^q)$, where q is the number of bits in a group, is a k_0-input, n_0-output finite-state machine of encoder memory order m_0 [9]. Thus, the set of k_0 data groups, each of a fixed length q, is input into an (n_0, k_0, m_0) convolutional encoder, and $(n_0 - k_0)$

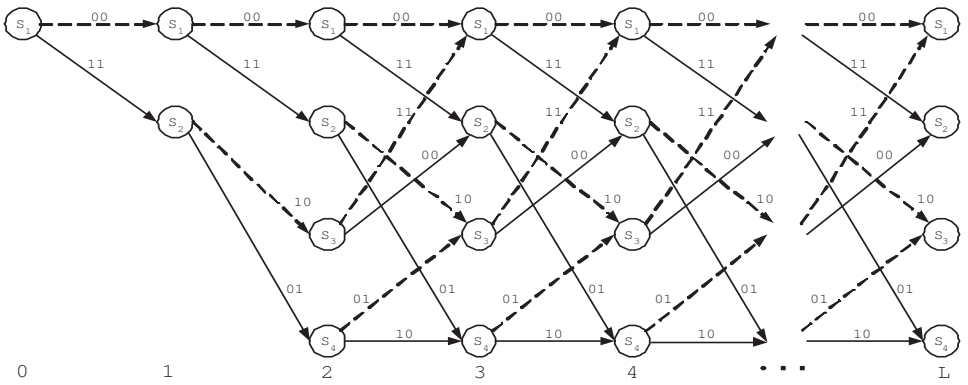

Figure 1: Trellis diagram for (2,1,2) convolutional code

redundant packets are generated based on a generator matrix. Parameter m_0 refers to the memory of the encoder, and indicates how many previous code groups influence the redundant packet.

In the process of encoding, information $m^{(u_i)}$ is firstly partitioned into the blocks that is introduced into the convolutional encoder to obtain the fingerprint codewords. And then these codewords are encoded by the inner encoder to the fingerprint sequences. Finally, these fingerprint sequences are concatenated into a fingerprint code. The $\Phi(L, N, l, n)$-code defined as follow: Suppose that a codeword $v = (v_1, v_2, \cdots v_L) \in \Im$ is an output of the convolutional encoder for information $m^{(u_i)}$, Let

$$W_v = (W^{(v_1)} \parallel W^{(v_2)} \parallel \cdots \parallel W^{(v_L)}) \tag{3}$$

where $\parallel$ means concatenation of strings. The code Φ is the set of all words W_v, i.e. $\Phi = \{W_v \mid v \in \Im\}$.

4.1.1 *fingerprint layer encoding*

The aim of the fingerprint layer is to resist collusion attacks at finite codebook and verify traitor codewords. Many codes can be regard as fingerprint layer, such as Identifiable Parent Property (IPP) codes, c-traceability (c-TA) codes, frameproof codes and so on. Note that the number c of traitors in a coalition cannot exceed the size n of codebook, namely, $n \geq c$. This paper adopts the (l, n) frameproof code Γ with c-secure to construct fingerprint layer. We choose the code $\Gamma_0(n, d)$ as an instance of (l, n)-code. The code $\Gamma_0(n, d)$ consists of all columns $(c_1, c_2, \cdots, c_{n-1})$ each duplicated d times. The amount of duplication determines the error probability ε. c_i each duplicated d times is called a Block that can be denoted as $B_i (1 \leq i \leq n-1)$. Let $\{w^{(1)}, w^{(2)}, \cdots, w^{(n)}\}$ denotes a codeword of $\Gamma_0(n, d)$, then the codeword $w^{(i)}$ is defined as follows:

$$w^{(i)} = \underbrace{\underbrace{00\cdots0}_{d}, \cdots, \underbrace{00\cdots0}_{d}}_{(i-1)d}, \underbrace{\underbrace{11\cdots1}_{d}, \cdots, \underbrace{11\cdots1}_{d}}_{(n-i)d} \tag{4}$$

where, $1 \leq i \leq n$ and the length of each fingerprint code is $d(n-1)$. When $i = 0$, the codeword is $w^{(0)} = \{1\}^{d(n-1)}$.

4.1.2 *Convolutional error-correcting layer encoding*

In $\Im(n_0, k_0, m_0)$, a user identification $m^{(u_i)}$ are divided into L groups with k_0 bits in each group, which involves m_0 groups to guarantee return to the initial state. $m^{(u_i)}$ is convolutional encoded as $v = (v_1, v_2, \cdots v_L)$ which is a binary sequence of length $n_0 L$. Here, $\Im$ is prone to choose the convolutional code with more free distance between codewords. We know that each v_i has $2^{k_0 m_0}$ states in state transition diagram. For the purpose of concatenating code, let $n \geq 2^{k_0 m_0}$ and each v_i is assigned to a codeword $W^{(v_i)}$ in Γ. Hence we can obtain L codes $(W^{(v_1)}, W^{(v_2)}, \cdots, W^{(v_L)})$ and concatenate those into a string. Finally, the string are randomly permuted by π to generate a codeword. The permutation π prevents the coalition from distinguishing the codeword $W^{(v_i)}$. Consequently, the distance restrict between codewords is eliminated in BS model. The security of fingerprinting depends only on the secret permutation π chosen randomly by merchant and the randomicity of the inner codes. The encoding process is described as follows: Suppose $X = (X_1, X_2, \cdots, X_L)$ is cover data from the original content S, where X_i is the sequence of length l, i.e. $X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,l}) \quad (i = 1, 2, \cdots, n)$. For user u_i , we can encode $m^{(u_i)}$ and number i as $v = (v_1, v_2, \cdots, v_L)$ by using convolutional encoder. And then each v_j is encoded $W^{(v_j)} = (w_1^{(v_j)}, w_2^{(v_j)}, \cdots, w_l^{(v_j)})$ in fingerprint larger. Furthermore, these codes are concatenated and randomly permuted by π to generate fingerprint code $W^{(v_j)}$, that is $W^{(v)} = \pi\left(W^{(v_1)} \parallel W^{(v_2)} \parallel \cdots \parallel W^{(v_L)}\right)$, where $\parallel$ can be realized in transform domain. Finally, a fingerprint $W^{(v)}$ is embedded into X to obtain the copy $X^{(u_i)}$ of user u_i. The outer encoding algorithm is described as follow:

Algorithm 1 (Convolutional fingerprint encoding algorithm)

```
Fingerprint-Information-Encoding(original-document X,
 message m^(u_i), permutation π)
    Let v = Convolutional_Encoding(m^(u_i), i)
    For each 1≤k≤L
        w^(v_k)=Fingerprint-Encoding(v_k)
    Let W^(v) = π(W^(v_1) ‖ W^(v_2) ‖ ··· ‖ W^(v_L))
    Embed W^v into document X to obtain X^(u_i)
    Return marked copy X^(u_i)
End
```

4.2 Convolutional fingerprinting decoding

The fingerprint decoding is a tracing algorithm that can identify traitors as many as possible by efficient means. It is requested that the tracing algorithm never accuses an innocent user and the probability that tracing fails can be made arbitrarily small. The identification algorithm of BS model involves the decoding of a random code, that is known to be a NP-hard problem. To resolve this problem, we present an optimal probability decoding algorithm by improving c-frameproof tracing algorithm and convolutional Viterbi algorithm.

4.2.1 fingerprint layer decoding

After the fingerprint is attacked by coalition, $\Gamma_0(l,d)$-decoder in BS model does not find out many illegal codewords but arbitrarily choose one of those codewords as a result of the limit of $(L,N,D)_p$-ECC codes. Moreover, the ultimate result of fingerprint decoding can only find at most a member of the guilty coalition. In contrast with BS model, we present a decoding algorithm of $\Gamma(l,n)$ code that can output a codeword set of coalition. Such a codeword set is called an Optional Code Set (OCS). Let $R_i = B_{i-1} \cup B_i$, similar to algorithm 1 in BS model, the tracing algorithm is improved as follows:

Algorithm 2 (fingerprint decoding algorithm)

```
Fingerprint-Layer-Decoding(fingerprinting-codeword x)
    Let U = {}
    If weight(x|_{B_1} > 0)
        Then U = U ∪ {1}
    For each 1≤k≤n−1
        Let s = weighgt(x|_{B_k})/2
        If weight(x|_{B_{k−1}}) < s − sqrt(s log(2n/ε))
            Then U = U ∪ {k}
    If weight(x|_{B_{n−1}}) < d
        Then U = U∪{n}
    Return U
End
```

4.2.2 convolutional error-correcting layer decoding

Maximum-likelihood (ML) decoding of convolutional codes is often implemented by means of the Viterbi algorithm. However, The Viterbi algorithm must be improved to perform optimal probability decoding because the codeword space of convolutional is extended by the optical code sets of inner codes. The Viterbi decoding is a minimum-distance probability decoding algorithm for convolutional codes. In trellis, assumption that the received symbol sequences is $R = (r_1, r_2, \cdots, r_L)$, where each optional code set r_i is composed by some suspicious codewords, i.e. $r_I = \{r_{i,1}, r_{i,2}, \cdots, r_{i,t}\}$ and $t \in N$. The optimal decoding tries to find out a shortest path that the encoder goes across in trellis, which is equivalent to compute a maximum-likelihood path among $2^{k_0 L}$ paths of length L, i.e. $\max_i(\log(\Pr(R|H_i)))$ $(1 \le i \le 2^{k_0 L})$, where $\Pr(R|H_i)$ is likelihood function between R and H_i. Let R_i be the set of all paths before stage i among the received sequence R, i.e. $R_i = (r_1, r_2, \cdots, r_i)$; C_i be all arrived branches at stage i; $C_{i,j}$ be the branches to arrive state S_j at stage i; $e_{i,j}$ be a branch from state S_i to S_j; function $D(X|Y)$ be the path metric between path X and Y. In stage i, there exist many paths $C_{i,j}$ to reach state S_j, but only the maximum-likelihood path among $C_{i,j}$ are called survivor path $sp_{i,j} = (e_{1,i_1}, e_{2,i_2}, \cdots, e_{i,i_i})$. the path metric between $sp_{i,j}$ and R_i is called part metric. Since maximum likelihood decoding and minimum distance decoding are the same for a Binary Symmetric Channel (BSC), the part metric has minimum hamming distance, i.e. $d_{i,j} = D(sp_{i,j}) = \min D(R_i|C_{i,j})$. Hence, we employs minimum distance to illustrate algorithms.

The proposed Viterbi algorithm can implement the maximum-likelihood decoding based on the optical code sets. The algorithm performs step-by-step as follows:

1. Initialization (at stage 0): Set the part metric of the original state S_1 of the trellis at 0 and others at ∞, the survivor path of each state is *null*.
2. Computation next stage: We suppose that at the previous stage k we have identified all survivor paths and stored each state's survivor path and part metric. For each state $S_j (1 \le j \le n)$ at stage $k+1$, the candidate path $sq_{i,j}$ is computed as the addition of all incoming branches $e_{i,j}$ and the survivor path $sp_{k,i}$ in connection with this branch, i.e. $sq_{i,j} = (sp_{k,i}, e_{i,j})$. In order to compute the minimum path metric $d_{k+1,j}$ between the candidate path $sq_{k+1,j}$ and the received path R_{k+1} before stage $k+1$ among R, for each incoming branch $e_{i,j}$, we compute the minimum metric $\min D(r_{k+1}|e_{i,j})$ between $e_{i,j}$ and optional code set r_{k+1}, and then the part metric $d_{k+1,j}$ of state S_j is computed as the minimum value of the addition of it and the part metric $d_{k,i}$ of state S_i, i.e.
$$\begin{aligned} d_{k+1,j} &= \min D(R_{k+1}|sq_{i,j}) \\ &= \min_i(d_{k,i} + \min D(r_{k+1}|e_{i,j})) \end{aligned} \quad (5)$$
where, the minimum metric between the optional code set $r_{k+1} = \{r_{k+1,1}, r_{k+1,2}, \cdots, r_{k+1,t}\}$ and the branch $e_{i,j}$ is computed by $\min D(r_{k+1}|e_{i,j}) = \min_{1 \le l \le m} D(r_{k+1,l}|e_{i,j})$. The path corresponding to $d_{k+1,j}$ is survivor path $sp_{k+1,j}$. Finally, we store each state's survivor path $sp_{k+1,j}$ and part metric $d_{k+1,j}$ and delete the candidate paths.
3. Final stage: If $k \le L$, then repeat step (2). Otherwise, we continue the computation until the algorithm reaches the termination symbol, at which time it makes a decision on the maximum-likelihood path that is equation to the survivor path corresponding to the minimum part metric $\min_{1 \le j \le n}(d_{L,j})$.

Finally, the decoding algorithm outputs the sequence of bits corresponding to this optimum path's branches. Convolutional fingerprint decoding algorithm is described at details in Algorithm (3).

5. PERFORMANCE ANALYSIS

theorem 3. The survivor path is maximum likelihood path in the improved Viterbi decoding algorithm (3) , namely, there exist survivor path sp for all candidate paths $sq \ne sp$, $D(R|sp) \ge D(R|sq)$.

PROOF. In trellis diagram of Viterbi decoder, let each surviving trellis path at all states is a maximum-likelihood paths before stage k. When the stage translates into $k+1$,

Algorithm 3 (Convolutional fingerprint decoding algorithm)

```
Convolutional-Decoding (suspect document Y, original document X)
    Let d_{0,i} = ∞ , sp_{0,i} = {} for (1≤i≤n) except d_{0,1} = 0
    Let W = π^{-1}(x)
    For each 1 ≤ k ≤ L
        Let r_k = Fingerprint-Layer-Decoding(Y_k , X_k )
        For each state s_j
            For each the incoming branch e_{i,j}
                For each the element r_{k,l}∈r_k (1≤l≤m)
                    c_{k,l} = D(r_{k,l} | e_{i,j} )
                Let sq_{i,j} = (sp_{k,i}, e_{i,j}), t_{i,j} = d_{k-1,i} + min_{1≤l≤m} c_{k,l}
            Let d_{k,j} = min_i(t_{i,j} ) for exist e_{i,j}
            Let sp_{k,j} = sq_{l,j} for all d_{k,j} == t_{l,j}
    Let d_{L,l} = min_{1≤j≤n}(d_{L,j} )
    Return M(sp_{L,l} )
End
```

according to the presented algorithm (3) the survivor path of S_j is $sp_{k+1,j} = (sp_{k,i}, e_{i,j})$, the corresponding part metric $d_{k+1,j}$ would satisfy the following relation:

$$\begin{aligned} d_{k+1,j} &= minD(R_{k+1}|sp_{k+1,j}) \\ &= d_{k,i} + minD(r_{k+1|e_{i,j}}), \end{aligned} \tag{6}$$

where, for all $r_{k+1,l} \in r_{k+1}$, $min\{D(r_{k+1})|e_{i,j}\} = min_l$ $D(r_{k+1,l}|e_{i,j})$. Using reduction to absurdity, assuming there exists a path is the different from $sp_{k+1,j}$ and its path metric is less than that of $sp_{k+1,j}$. Let this path is $sp'_{k+1,j} = (p_{k,i'}, e_{i',j})$ and $i \neq i'$, where $p_{k,i'}$ denotes k-edges before stage k. Thus the part metric of $sp'_{k+1,j}$ is computed by

$$\begin{aligned} d'_{k+1,j} &= minD(R|sp'_{k+1,j}) \\ &= minD(R_k|p_{k,i'}) + minD(r_{k+1}|e_{i',j}) \\ &= d'_{k,i'} + minD(r_{k+1}|e_{i',j}), \end{aligned} \tag{7}$$

where $d'_{k+1,j} = minD(R_k, p_{p,i'})$. Therefore, since $sp_{k+1,j}$ is a survivor path, the path $sp_{k+1,j}$ has the least metric among all paths reaching state S_j at stage $k+1$. That is, for all $1 \leq l \leq n$ and $i \neq l$, there holds

$$\begin{aligned} minD(R_{k+1}|sp_{k+1,j}) &= d_{k,i} + minD(r_{k+1}|e_{i,j}) \\ &\leq minD(R_{k+1}|(C_{k+1}, e_{l,j})) \\ &= d_{k,l} + minD(r_{k+1}|e_{l,j}). \end{aligned} \tag{8}$$

For i', if $i \neq i'$, then it is clear from (8) that

$$d_{k,i} + minD(r_{k+1}|e_{i,j}) \leq d_{k,i'} + minD(r_{k+1}|e_{l,j}). \tag{9}$$

Since the metric of $sp'_{k+1,j}$ is less than that of $sp_{k+1,j}$ ($d'_{k+1,j} < d_{k+1,j}$), by (6),(7) and (9) we have

$$\begin{aligned} d'_{k+1,j} &= d'_{k,i'} + minD(r_{k+1}|e_{i',j}) \\ &< d_{k+1,j} = d_{k,i} + minD(r_{k+1}|e_{i,j}) \\ &\leq d_{k,i'} + minD(r_{k+1}|e_{i',j}). \end{aligned} \tag{10}$$

We can know $d'_{k,i'} < d_{k,i'}$. However, this is contradictions with the assumption that the state $S_{i'}$ has maximum-likelihood path at stage k. This proves the theorem. □

For the purpose of protecting innocent user from needless accusation, according to the properties of c-frameproof code and convolutional code, we can prove that the presented algorithm can find a member of the coalition with probability at least $1 - \varepsilon$.

theorem 4. Given integers N, c and $\varepsilon > 0$, set $n = 2c$, $d = 2n^2(\log(8n)+r)$, $r = (2/d_f)\log(A_{d_f}/\varepsilon)$, where d_f is free distance of the code, A_{d_f} is the number of the code with weight d_f. Then convolutional fingerprinting code $\Phi(L, N, n, d)$ is a code which is c-secure with ε-error. The code contains N codewords. Let x be a word which was produced by a coalition C of at most c users. Then Algorithm (3) will output a codeword of C with probability at least $1 - \varepsilon$.

Proof. According to the properties of convolutional codes, in BSC channel, the error probability P_e of Viterbi decoder is

$$P_e \approx A_{d_f} 2^{d_f} p^{d_f/2} \tag{11}$$

where, d_f is free distance of the code, A_{d_f} is the number of the code with weight d_f and p is channel transfer probability. Let the decoding error probability of frameproof code $\Gamma(l, n)$ is ε', the channel transfer probability is equation to the error probability of fingerprint codes Γ, i.e. $p = \varepsilon' = \frac{1}{4}(P_e/A_{d_f})^{2/d_f}$. On the basis of collusion-resistant properties of the fingerprint code Γ in Theorem (2), the codeword length is $d = 2n^2 log(2n/\varepsilon')$. Then, the codeword length of Φ is

$$d = 2n^2(\log(8n) + (2/d_f)\log(A_{d_f}/P_e)) \tag{12}$$

Notice that the properties of convolutional codes decides that the codeword length be independent of the error probability. Since $P_e = \varepsilon$ in Φ, the code Φ may occur the decoding error with probability at most ε when x be a fingerprint word which was produced by a coalition C of at most c users. Moreover, Theorem (4) denotes that the improved decoding algorithm can perform effectively the maximum likelihood search. As a result, algorithm (3) will output a codeword of C with probability at least $1 - \varepsilon$. □

Here, we don't intent to discuss the performance of encoding algorithm because it is obvious that the algorithm is lower complexity than previous algorithms. we focus our attention on the decoding algorithm from the following aspects:

Encoding length The code length of fingerprinting encoder is $Ld(n-1)$. At the same time, in BS model each $\Gamma_0(l, d)$-code bears with the error probability of $\varepsilon/2L$ and L depend on N and ε. Hence block length d is augmented along with increases of L. But in the proposed scheme d increase fix $2rn^2 = 8rc^2$ bits in terms

of Theorem (5.2) since the current group depends only on m_0 previous groups according to the property of convolutional codes. As a result, L is independent of d, the code length is shorter than the other FP codes and then the parameters of convolutional fingerprinting code can be predefined.

Decoding complexity The decoding complexity of frameproof codes is in direct ratio to $O(nLl)$ and (n_0, k_0, m_0) convolutional codes is $O(nL)$ by processing L steps and searching n states at each step. The whole decoding complexity is still $O(nLl)$.

Storage performance In respect of storage, general convolutional decoder must retain $n = 2^{k_0 m_0}$ states and only the survivor path and its metric must be stored for each state at the current trellis stage, as the decoding algorithm progresses. Hence it is obvious that the storage complexity of the whole decoder is $O(nL)$, where $n = 2^{k_0 m_0}$ and $n \geq c$. Usually $m_0 \leq 10$ and the length of the user information $m^{(u)}$ requests $|m| = k_0 L$. However, for finite length of the original medium, the proposed code allows us to shorten L by adjusting k_0 and m_0. It isn't difficult to choose the better convolutional code even if the coalition size c is larger.

At present, most of fingerprinting schemes are constructed by using BS model, such as Identifiable Parent Property code (IPP) [10], c-Traceability code (c-TA), Frameproof code (FP) [11] and so on. Although they have the different methods for encoding and decoding, the original properties of BS model are kept by using concatenation structure, and then there exist some questions about encoding length and decoding performance [12]. However, the proposed scheme has been close to the lower bounds for collusion-secure fingerprinting. Moreover, the traitor searching is also implemented with finite resources [13].

6. APPLICATION EXAMPLE

We refer to a small example extracted from the process of fingerprint scheme in order to illustrate the construction methods and search strategies presented in this paper. The example given here is deliberately simple. Consider a $(2, 1, 2)$ convolution encoder with code generators $g^{(1)}(D) = 1 + D^2$ and $g^{(2)}(D) = 1 + D + D^2$. The free distance d_f of this convolution code is 5 and the sequence number A_{d_f} of weight-5 is 1. The encoder register has length $m = 2$, the number of state is 4. The state diagram for encoder is shown in Fig.2, where the solid line corresponds to message 1 and the dashed line to 0.

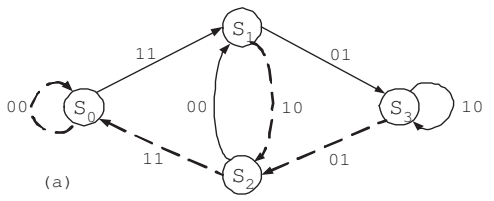

Figure 2: state diagram for (2,1,2) convolutional code

Here the fingerprint code employs the code $\Gamma = \Gamma_0(4, d) = \{111, 011, 001, 000\}_d$, in which each bit is extended d-times.

Table 1: Encoding table for $\Gamma_0(4, 4)$ code

codeword	1st block	2rd block	3th block
w^1	1111	1111	1111
w^2	0000	1111	1111
w^3	0000	0000	1111
w^4	0000	0000	0000

For example, the code $\Gamma_0(4, 4)$ is shown in Table 1 for $d = 4$. According to Theorem 5.2, we knows that this code enables us to resist $c = 2$ collusion for the users $n = 4$, error probability $\varepsilon = 0.0001$ and extending $d \geq 300$. We assume without less of generality that a codeword of the encoding table in Table 1 is defined as a letter in $\Sigma = \{1, 2, 3, 4\}$, e.g. w^1 corresponds to 1, and the information length $L = 7$, the encoding and decoding process are illustrated for the presented scheme as follows.

Suppose that two user u_1 and u_2 choose at random the identification information $M_1 = (1011100)$ and $M_2 = (0110100)$, respectively, where the last two bits is the ending symbol. Let us first describe encoding process: at first, the codewords corresponding to the information M_1 and M_2 are obtained by the encoder as above, namely, the output of encoder is the sequences $R^{(1)} = (11, 10, 00, 01, 10, 01, 11)$ and $R^{(2)} = (00, 11, 01, 01, 00, 10, 11)$. Secondly, assume that the alphabet $\Sigma = \{1, 2, 3, 4\}$ are defined by the encoding table in Table 1, and then in the codeword sequence let every two bits correspond to a letter symbol in Σ, namely, $R^{(1)}$ and $R^{(2)}$ are encoded to $R^1 = (4, 3, 1, 2, 3, 2, 4)$ and $(R^2 = (1, 4, 2, 2, 1, 3, 4)$, respectively. Thirdly, the fingerprint codeword $W^{(1)}$ and $W^{(2)}$ are obtained by concatenating the sequences in the Γ_0 code, namely, $W^{(1)} = (w^4, w^3, w^1, w^2, w^3, w^2, w^4)$ and $W^{(2)} = (w^1, w^4, w^2, w^2, w^1, w^3, w^4)$, where w^i $(1 \leq i \leq 4)$ is shown as Table 1. Finally, the merchant randomly chooses a permutation π and then computes $\pi(W^{(1)})$ and $\pi(W^{(2)})$ to construct the fingerprinting codewords.

Assume that the coalition $C = \{u_1, u_2\}$ produces an illegal copy with a fingerprint $y \in \Sigma^L$. The merchant expects to trace at least one of its member after he captures this illegal copy. To accomplish this, at first, the marks are extracted from the copy, and then the merchant performs the reverse permutation π^{-1} on the mark sequence to obtain a fingerprint mark sequence W', where the ordering is determined by secret key. Secondly, according to algorithm (2), the optional code set $r_i (1 \leq i \leq L)$ could be produced from each block in mark sequence, note that r_i is likely to involve more than one letter. A fingerprint sequence R is formed by concatenating all r_i. Here we assume that the fingerprint sequence extracted is $R = (\{w^4\}, \{w^3, w^4\}, \{w^1, w^2\}, \{w^2\}, \{w^1\}, \{w^2, w^3\}, \{w^4\})$ that is arbitrarily manipulated by colluders according as marking assumption. Thirdly, R is translated into the binary sequence R' by Γ encoding table, namely $R' = (\{11\}, \{10, 11\}, \{00, 01\}, \{01\}, \{00\}, \{01, 10\}, \{11\})$. And then R' is putted into convolutional decoder in algorithm (3). The process of decoding shown in Fig.3 is composed of 7 steps. Note that the most difference is that the proposed decoding algorithm can deal with optional code set

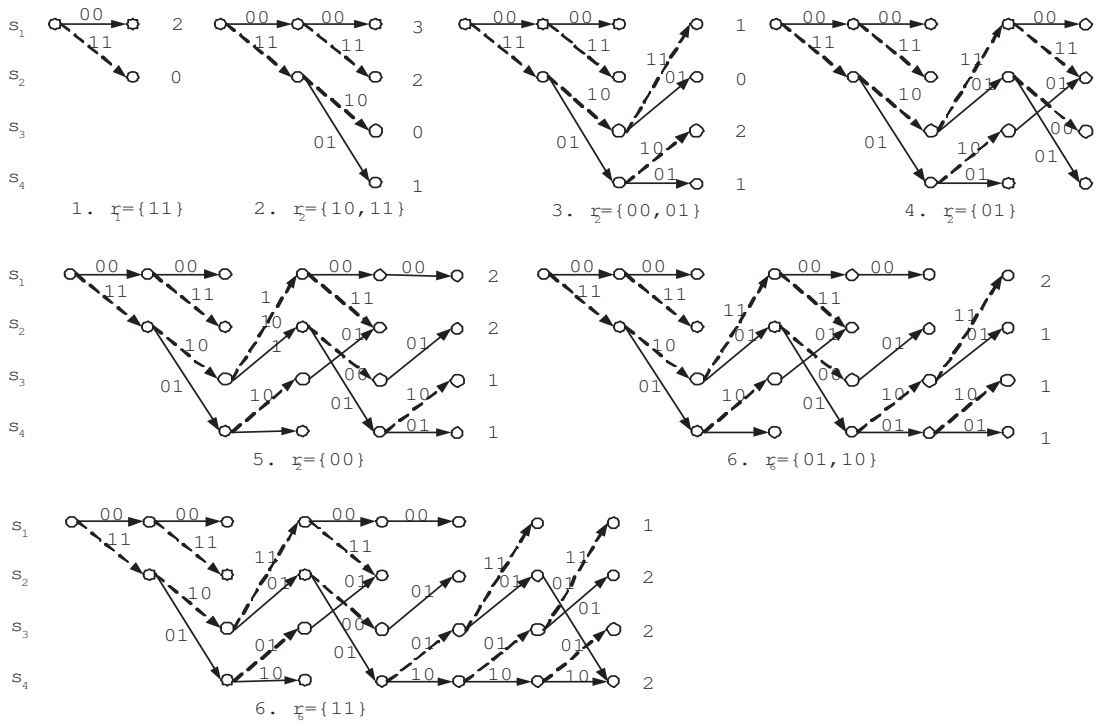

Figure 3: Trellis diagrams of the decoding process for the improved Viterbi algorithm

r_i instead of a single letter. For example, in the step 3, the input is $r_3 = \{00, 01\}$, the state S_1 has two entering branch $e_{1,1}$ and $e_{3,1}$. For $e_{1,1}$, the minimum distance between $e_{1,1}$ and r_3 is $M(r_3|e_{1,1}) = min(D(00|00), D(01|00)) = 0$ and its candidate path metric is $d'_{2,1} + M(r_3|e_{1,1}) = 3$. For $e_{3,1}$, the minimum distance between $e_{3,1}$ and r_3 is $M(r_3|e_{3,1}) = minD(00|11), D(01, 11) = 1$ and its candidate path metric is $d'_{2,3} + M(r_3|e_{3,1}) = 1$. Hence, the result of the algorithm is that the survive path passes through $e_{3,1}$ and the minimum distance is $d_{3,1} = 1$.

Finally, after the algorithm executes 7 steps decoding, the survivor path is $(11, 10, 00, 01, 10, 01, 11)$, the minimum Hamming distance between this path and the input R' is 1. The user information correspond to it is $M_1 = (1011100)$. This result indicates that the user u_1 is one number of the colluding group. In conclusion, the proposed fingerprinting scheme can provide cost-effective protection against collusion and rapidly implements the efficient Traitor Tracing.

7. CONCLUSION

This paper provides a new approach for encoding and decoding fingerprint for the user private information based on convolutional code with optional code set. This encoding system is easy to implement and has acceptably low complexity. Furthermore, It also has significant reference value and guidance meaning for Intellectual Property Protection and relative field in theory and practice.

8. ADDITIONALAUTHORS

Additional authors: Yang Yongtian (Computer Science and Technology School, Harbin Engineering University, email: yangyt8@hotmail.com) and Feng Dengguo (The State Key Laboratory of Information Security, Chinese Academy of Sciences, email: fengdg@263.net).

9. REFERENCES

[1] D. Boneh and J. Shaw. Collusion-Secure Fingerprinting for Digital Data. In Advances in Cryptology - CRYPTO 95, Lecture Notes in Computer Science, Berlin: Springer-Verlag, 1995, 963:452-465.

[2] A.Barg, G.R. Blakly, and G. Kabatiansky. Digital Fingerprinting Codes: Problem Statements, Constructions, Identification of Traitors. Technical report, DIMACS2001-52, 2001

[3] Y. Wang, S.-W. Lu, H.-L. Xu. A Digital Fingerprinting Algorithm Based on Binary Codes. Journal of software, 2003,14(06): 1172-1177. (in Chinese)

[4] F. Ergun, J. Kilian, and R. Kumar, A note on the limits of collusion-resistant watermarks, in Eurocrypt '99, Lecture Notes in Computer Science **1592**, Berlin: Springer-Verlag, 1999: 140–149

[5] I. Biehl and B. Meyer. Protocols for Collusion-Secure Asymmetric Fingerprinting (Extended Abstract), Proceedings of the 14th Annual Symposium on Theoretical Aspects of Computer Science. 1997: 399-412

[6] M. Fernandez and M. Soriano. Identification of Traitors in Algebraic-Geometric Traceability Codes. in

IEEE Trans. on Signal Processing. Supplement on Secure Media , 2004,52(10): 3073–3077

[7] J. Kilian, F. T. Leighton, L. R. Matheson, T. G. Shamoon, R. E. Tarjan, and F. Zane. Resistance of Digital Watermarks to Collusive Attacks. Technical Report TR-585-98, Princeton University, Computer Science Department, July 1998. http://citeseer.ist.psu.edu/ kilian98 resistance.html

[8] F. Chan and D. Haccoun. Adaptive Viterbi Decoding of Convolutional Codes over Memoryless Channels. IEEE Transactions on Communications, 1997,45(11): 1389 -1400.

[9] GD Forney Jr. Convolutional Codes I: Algebraic Structure, IEEE Trans. on Information Theory, **IT-16**(6), November, 1970: 720–738

[10] G. Cohen , S. Encheva, S. Litsyn. Intersecting codes and partially identifying codes. In International Workshop on Coding and Cryptography. Paris: Elsevier Press, 2001:139-147.

[11] J.N. Staddon , D.R. Stinson , R. Wei. Combinatorial properties and constructions of traceability schemes and flameproof codes. SIAM Journal on Discrete Math, 1998, 11(1):41-53.

[12] V. Guruswami and M. Sudan. Improved decoding of reed-solomon and algebraic-geometry codes. IEEE Trans. on Information Theory, 1999, 45(6):1757-1767.

[13] Peikert C, Shelat A, Smith A. Lower bounds for collusion-secure fingerprinting. In 14th Annual ACM-SIAM Symposium on Discrete Algorithms, Edmonton: ACM Press, 2003:472-479.