A.L. Varna and M. Wu, "Theoretical Modeling and Analysis of Content Identification", invited paper in Special Session on Content Identification/Fingerprints, IEEE International Conference on Multimedia and Expo (ICME), July 2009.

MODELING AND ANALYSIS OF CONTENT IDENTIFICATION

Avinash L. Varna and Min Wu

Dept. of ECE and Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA.

ABSTRACT

Content fingerprinting provides a compact content-based representation of a multimedia document. An important application of fingerprinting is the identification of modified copies of the original media content. These modifications may be incidental changes that occur during the usage of multimedia, or intentional modifications made by an adversary to avoid detection. Currently, the effectiveness of content identification techniques is often assessed through benchmark databases. To complement these experimental performance evaluations, this paper develops a theoretical framework for analyzing content identification techniques. Beneficial aspects from decision theory and game theory are exploited to gain insights toward optimal system design and parameter selection.

Index Terms— Content identification, decision theory, game theory, Nash equilibrium.

1. INTRODUCTION

Content fingerprinting provides a compact content-based representation of a multimedia document. An important application of fingerprinting is the identification of modified copies of the original media. These modifications may be incidental changes that occur during the usage of multimedia, or intentional modifications made by an adversary to avoid detection. Content fingerprinting has found applications in content filtering on user-generated content (UGC) websites, such as YouTube. The fingerprint of each video uploaded to the website is compared against a database of copyrighted content. If a match is found, the owner may be notified and provided with different options, such as removing the content, or sharing the revenue generated from the video. Fingerprinting has also found applications in associating unannotated data to the appropriate metadata and creating services that allow users to identify music by recording short clips on their cell phones.

Several fingerprinting techniques exist in the literature for performing multimedia content identification, and are mostly evaluated using benchmark databases. From these evaluations, it is difficult to infer how the performance would scale when the system is used in a practical application involving large databases with millions of multimedia documents. Theoretical modeling and analysis can help us understand how the performance of such schemes would scale as the number of reference content becomes large. A decision theoretic framework for analyzing content identification systems that utilize binary hashes was proposed in [1]. The minimum length of a hash required to achieve a desired low probability of false alarm and a high probability of detection when the size of the database is of the order of a billion was derived.

The analysis in [1] was performed from the system designer's perspective. To complement these previous results, in this paper, we model the interaction between the system designer and a malicious adversary attempting to subvert the system under a game theoretic framework. This analysis helps us understand the effect of different distortions a video may undergo and suggests strategies for designing fingerprints to achieve the best possible performance.

In the content identification problem, the system designer and the adversary modifying the content have conflicting objectives. The adversary's goal is to avoid detection, while the designer's goal is to identify modified content and minimize the probability of misclassification and false alarm. In this paper, we model the dynamics between the designer and the adversary by a two-player game. We consider the use of binary hashes for content identification, and derive optimal strategies for the system designer and the adversary to maximize their respective utilities.

This paper is organized as follows: Section 2 reviews the main results from a decision-theoretic analysis of content identification presented in [1]. Section 3 describes the game theoretic model using binary hashes as an example and derives optimal strategies for the adversary and the system designer. Section 4 summarizes the results and contribution of this paper.

2. DECISION-THEORETIC ANALYSIS OF CONTENT IDENTIFICATION

Our recent research carried out a decision-theoretic analysis of content identification to understand how the probability of correct identification and false positives scales as a function of the database size, when the number of reference videos is very large [1]. In this section, we review the analysis and main results.

Content identification was modeled as a multiple hypothesis testing problem, where given the fingerprint/hash of a video under question, the detector aims to determine whether

Email:{varna,minwu}@eng.umd.edu

the query video is a (possibly modified) version of some video in the database, and if so, identify the original video. Given a reference database of N videos $\{V_1, V_2, \ldots, V_N\}$ and the query video Z, the detector performs the following multiple hypothesis test:

$$\begin{aligned} H_0 &: Z \text{ is not from the database } \{V_1, V_2, \ldots, V_N\}, \\ H_1 &: Z \text{ corresponds to video } V_1, \\ &\vdots \\ H_N &: Z \text{ corresponds to video } V_N. \end{aligned} \quad (1)$$

The detector computes the fingerprint $\mathbf{y}$ of the query Z and compares it with the fingerprints $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of the videos in the database. We considered the scenario of using binary fingerprints for identification in [1], assuming that the individual bits of the fingerprints are independent and equally likely to be 0 or 1. The possible modifications to the content were characterized by the average fraction of hash bits p that were altered as a result of the modifications. Under this setting, the probability of correctly identifying the content and the probability of false positives was determined as a function of the parameter p, when the number of videos N is of the order of a billion.

From a system designer's perspective, guidelines were provided for choosing the length of the hash to achieve a desired performance in terms of the probability of detection P_d and false positives P_f. Given a database of N videos, to achieve $P_d \approx 1$ such that $P_f \leq \epsilon$, it was found that the length of the hash L must satisfy

$$\frac{1}{L} \log_2 \frac{N}{\epsilon} < 1 - h(p), \quad (2)$$

where p is the average fraction of hash bits that are changed when the original content is modified and $h(p) = -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. Experimental results were presented in [1] and agree well with the theoretical predictions.

The above analysis was performed from a system designer's perspective and provides guidelines for choosing the system parameters to achieve a desired performance. The effect of possible modifications to the content was represented by the average fraction of fingerprint bits that are altered as a result of the modification. In the subsequent sections we look more closely at the other side of the coin. We examine the dynamics of the interaction between the designer and an adversary seeking to subvert the system and suggest strategies for designing the fingerprints to achieve the best possible performance.

3. GAME-THEORETIC ANALYSIS OF CONTENT IDENTIFICATION

In this section we model content identification using content fingerprinting/hashes as a game between two players, the adversary $\mathcal{A}$ and the system designer $\mathcal{D}$. In this game, the designer $\mathcal{D}$ designs the hashing scheme and the adversary chooses the attack to maximize their respective payoff functions. We illustrate this model using the example of binary hashes, which are commonly used for content identification [2–4].

3.1. Strategy Space

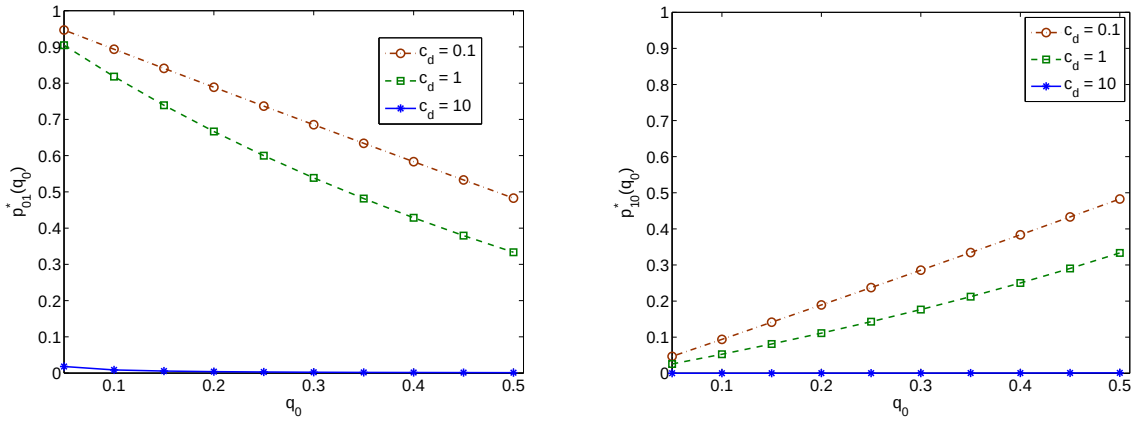
In the content identification game using binary fingerprinting, the strategy space of the designer consists of possible choices for the distribution of the hash bits. For simplicity, here we consider hash bits that are independent and identically distributed (i.i.d.). Under this setting, the designer chooses a value $0 \leq q_0 \leq 0.5$ as the probability that a hash bit is 0 and $q_1 = 1 - q_0$ is the probability that a hash bit is 1. Thus, the strategy space for the designer S_D is the interval $[0, 0.5]$.

The strategy space for the adversary consists of possible modifications of the content that do not introduce excessive distortion and render the content unusable. Denote the probability of a hash bit 0 being changed to a 1 after modification of the video by p_{01} and the probability that a bit 1 changes to 0 by p_{10}. As the adversary chooses these parameters, his strategy space is given by $S_A = 0 \leq p_{01}, p_{10} \leq 1$.

3.2. Designer's Payoff Function

At the detection stage, for each content V_i in the database, the detector has to decide whether the query content denoted by Z is a distorted version of V_i, by comparing their hashes. Let $\mathbf{x}_i$ and $\mathbf{y}$ be the hash of V_i and Z, respectively. If Z is indeed a modified version of V_i, then the hashes $\mathbf{x}_i$ and $\mathbf{y}$ are dependent and their joint distribution is $p(\mathbf{y}|\mathbf{x}_i)q(\mathbf{x}_i)$, where $q(\mathbf{x})$ is the distribution of the hashes and $p(\mathbf{y}|\mathbf{x})$ is the conditional distribution representing the modification. If Z is *not* a modified version of V_i, then the hashes $\mathbf{x}_i$ and $\mathbf{y}$ are independent and their joint distribution is $q(\mathbf{y})q(\mathbf{x}_i)$.

The system's performance can be characterized by the probability P_f of incorrectly deciding that $\mathbf{x}_i$ and $\mathbf{y}$ are dependent when they are actually independent (false positive), and the probability P_m of deciding that $\mathbf{x}_i$ and $\mathbf{y}$ are independent when they are actually dependent (missed detection). As the designer's objective is to achieve low values for P_f and P_m, a suitable function of P_f and P_m can be chosen as the payoff for the designer. However, as in any detection problem, these error probabilities are not independent of each other. In many practical applications, it is common to fix one of these error probabilities, say P_f, to be less than a threshold α and then minimize the other type of error. From the Chernoff-Stein Lemma [5], we know that the best asymptotic error exponent that can be achieved under this setting is given by the Kullback-Leibler (KL) distance between the distributions under the two hypotheses $D(p(\mathbf{y}|\mathbf{x}_i)q(\mathbf{x}_i)||q(\mathbf{y})q(\mathbf{x}_i))$. As the hash bits are i.i.d., the KL distance between the distributions is LD_{KL}, where $D_{KL} = D(p(y|x)q(x)||q(y)q(x))$, $p(\cdot|\cdot)$ is the conditional distribution representing the modification of one bit and $q(\cdot)$ is the common distribution of the



(a) $p^*_{01}(q_0)$ (b) $p^*_{10}(q_0)$

Fig. 1. Optimum choices of (a) p_{01} and (b) p_{10} for the adversary as a function of the system designer's choice of q_0.

individual hash bits. By choosing q_0 appropriately to maximize the KL distance, the designer can reduce the probability of making an error. Thus, we choose the KL distance between the two distributions as the payoff (utility) for the designer $U_D(q_0, p) = D_{KL} = D(p(y|x)q(x)||q(y)q(x))$.

3.3. Adversary's Payoff Function

The adversary's main goal is to evade detection while minimizing the amount of distortion introduced into the content. By choosing the parameters p_{01} and p_{10} to minimize the KL distance D_{KL} between the two distributions, the adversary can reduce the probability of being detected. Hence, we choose $-D_{KL}$ as the adversary's payoff function. We also add a penalty term to the adversary's payoff based on the amount of distortion introduced into the video, to incorporate the adversary's goal of minimizing the perceptual distortion. We assume that the distortion of the original video can be equivalently represented in terms of the change in the fingerprint of the video. For simplicity, we assume that the perceived commercial value of the distorted content reduces as a linear function of the Hamming distance between the hash of the original and modified content. The analysis can be performed similarly for other models relating the distortion to the reduction in commercial value.

Under this setting, the expected utility for the adversary can be given as $U_A(q_0, p) = -D_{KL} - c_d \frac{1}{L} E[d_H(\mathbf{x}, \mathbf{y})]$, where $E[d_H(\mathbf{x}, \mathbf{y})]$ is the expected Hamming distance between the hash $\mathbf{x}$ of the original content and the hash $\mathbf{y}$ of the distorted content, and c_d is the rate at which the perceived value of the content reduces as a function of the average Hamming distance. Since the hash bits are i.i.d., $\frac{1}{L} E[d_H(\mathbf{x}, \mathbf{y})] = q_0 p_{01} + q_1 p_{10}$ and the expected payoff for the adversary is $U_A(q_0, p) = -D_{KL} - c_d(q_0 p_{01} + q_1 p_{10})$. We see that the adversary can reduce the probability of getting caught by reducing D_{KL}, but this would increase the distortion and hence reduce the value of the content. The adversary has to find the optimal tradeoff between these conflicting objectives.

3.4. Subgame Perfect Equilibria

We recognize that under the above settings, the game corresponds to a sequential two player game with perfect recall [6]. In such sequential games, the optimal strategies for the players are given by Subgame Perfect Nash Equilibria (SPNE). The SPNE are similar to saddle-points and correspond to strategies from which neither player has incentive to deviate, given that the other player plays his equilibrium strategy. In other words, given that the designer plays his part of the equilibrium solution, the adversary cannot obtain a higher payoff by playing any strategy other than his equilibrium strategy, and vice versa. The SPNE of this game are given by points $(q_0^*, p^*(q_0^*))$, such that

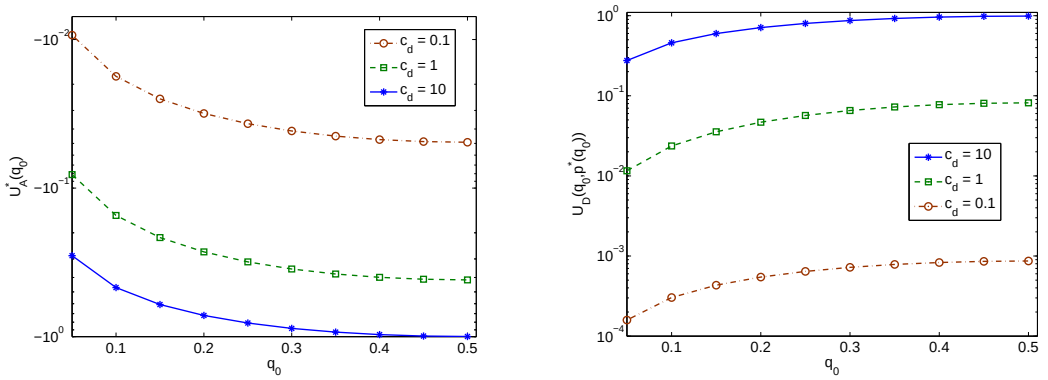
$$\begin{aligned} p^*(q_0) &= \arg \max_{0 \le (p_{01}, p_{10}) \le 1} U_A(q_0, p) \\ q_0^* &= \max_{0 \le q_0 \le 0.5} U_D(q_0, p^*(q_0)). \end{aligned} \tag{3}$$

The maximum expected payoff that the adversary can achieve, given that the designer chooses q_0 is given by

$$U_A^*(q_0) = \max_{0 \le p_{01}, p_{10} \le 1} -D_{KL} - c_d(q_0 p_{01} + q_1 p_{10}).$$

As $-D_{KL}$ is concave in p [5], the utility function U_A is concave in p. As the constraints are also concave, there is a unique maximizer which is determined as $p^*_{01}(q_0) = \frac{q_1 2^{-c_d}}{q_0 + q_1 2^{-c_d}}$ and $p^*_{10}(q_0) = \frac{q_0 2^{-c_d}}{q_1 + q_0 2^{-c_d}}$. Using the above values for p^*_{01} and p^*_{10}, the maximum value of the expected utility for the adversary is found to be $U_A^*(q_0) = q_0 \log_2(q_0 + q_1 2^{-c_d}) + q_1 \log_2(q_1 + q_0 2^{-c_d})$.

Fig. 1 shows the optimal values for p_{01} and p_{10} as a function of q_0 for various values of the degradation parameter c_d. We observe that when c_d is small, e.g. $c_d = 0.1$, which implies that the value of the distorted content reduces slowly as a function of the distortion introduced, the adversary can choose large values for p_{01} and p_{10}, corresponding to making large changes to the content so as to evade detection, without



(a) $U_A^*(q_0)$ (b) $U_D(q_0, p^*(q_0))$

Fig. 2. (a) Maximum payoff for adversary as a function of q_0. (b) Designer's payoff when the adversary plays his best strategy.

incurring a significant reduction in the commercial value. If the parameter c_d is large, e.g. $c_d = 10$, the adversary cannot introduce much distortion into the content, as the value reduces rapidly and is restricted to modifications that result in a very small fraction of the hash bits being altered. The maximum payoff that the adversary can obtain by playing his optimal strategy, in response to the designer's choice of q_0 is shown in Fig. 2(a). For any fixed value of q_0, the adversary obtains a higher payoff when c_d is small, as he can introduce distortion without reducing the value of the content significantly.

When the adversary plays his best response strategy $p^*(q_0)$ shown in Fig. 1, the payoff for the designer is found to be

$$\begin{aligned} U_D(q_0, p^*(q_0)) \quad = \quad & -q_0 \log_2(q_0 + q_1 2^{-c_d}) \\ & -q_1 \log_2(q_1 + q_0 2^{-c_d}) \\ & -c_d q_0 q_1 \frac{1 + 2^{-c_d}}{(q_0 + q_1 2^{-c_d})(q_1 + q_0 2^{-c_d})}, \end{aligned}$$

and is shown in Fig. 2(b). We observe that when c_d increases, the designer can obtain a higher payoff, as the adversary can make limited changes to the content. This indicates that the hash function should be designed carefully, so that it is not easy to alter hash bits without causing a lot of distortion. From the figure, we also see that for a fixed c_d, the payoff for the designer is an increasing function of q_0, attaining a maximum at $q_0 = 0.5$. Thus, the optimal strategy for the designer is to choose the hash bits to be 0 or 1 with equal probability, while the corresponding best strategy for the adversary is $p_{01} = p_{10} = \frac{1}{1+2^{c_d}}$. If $2^{c_d} \gg 1$, $p_{01} = p_{10} \approx 2^{-c_d}$ indicating that the optimal choice for the adversary is to modify a very small fraction of the bits. If $2^{c_d} \ll 1$, then $p_{01} = p_{10} \approx 1$ signifying that the adversary can cause large changes to the hash and easily evade detection.

4. CONCLUSIONS

In this paper, we have developed a theoretical framework to analyze content fingerprinting and identification. Using ideas from detection theory, we have examined the performance of identification systems as the number of reference content becomes very large and provided guidelines for choosing the length of the fingerprint to achieve a suitably low probability of error while maintaining a high probability of detection.

We have also modeled the dynamics of the interaction between the fingerprint system designer and an adversary seeking to subvert the system under the framework of game theory. Using the example of binary fingerprint-based content identification, we have illustrated our model and suggested strategies for designing the fingerprints to achieve the best possible performance. We showed that the optimal strategy for the system designer is to design the fingerprinting scheme such that the resultant fingerprint bits are equally likely and also highlighted the benefit of designing robust fingerprint schemes such that the content has to be distorted significantly in order to cause changes to the fingerprint.

5. REFERENCES

[1] A. L. Varna, A. Swaminathan, and M. Wu, "A Decision-Theoretic Framework for Analyzing Binary Hash-based Content Identification Systems," in *Proc. of the ACM Workshop on Digital Rights Management*, Alexandria, VA, Oct. 2008, pp. 67–76.

[2] J. Haitsma, T. Kalker, and J. Oostveen, "Robust Audio Hashing for Content Identification," in *International Workshop on Content-Based Multimedia Indexing*, Brescia, Italy, Sept. 2001.

[3] B. Coskun, B. Sankur, and N. Memon, "Spatio-temporal Transform Based Video Hashing," *IEEE Trans. on Multimedia*, vol. 8, no. 6, pp. 1190–1208, Dec. 2006.

[4] S. Baluja and M. Covell, "Content Fingerprinting using Wavelets," in *Proc. of IET Conference on Multimedia*, London, England, Nov. 2006.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, second edition, 2004.

[6] M. Osborne and A. Rubinstein, *A Course in Game Theory*, MIT Press, first edition, 2001.