A Decision Theoretic Framework for Analyzing Binary Hash-based Content Identification Systems

Avinash L. Varna
Department of Electrical and
Computer Engineering
University of Maryland
College Park, MD, USA
varna@umd.edu

Ashwin Swaminathan
Department of Electrical and
Computer Engineering
University of Maryland
College Park, MD, USA
ashwins@umd.edu

Min Wu
Department of Electrical and
Computer Engineering
University of Maryland
College Park, MD, USA
minwu@umd.edu

ABSTRACT

Content identification has many applications, ranging from preventing illegal sharing of copyrighted content on video sharing websites, to automatic identification and tagging of content. Several content identification techniques based on watermarking or robust hashes have been proposed in the literature, but they have mostly been evaluated through experiments. This paper analyzes binary hash-based content identification schemes under a decision theoretic framework and presents a lower bound on the length of the hash required to correctly identify multimedia content that may have undergone modifications. A practical scheme for content identification is evaluated under the proposed framework. The results obtained through experiments agree very well with the performance suggested by the theoretical analysis.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Security

Keywords

Content identification, content fingerprinting, decision theory

1. INTRODUCTION

Websites such as Youtube [1] have revolutionized content sharing services by making it easy for users to upload and share video. At the same time, concerns have been raised regarding potential copyright violations, as users may upload copyrighted content to these websites [2]. To counter such activity, content identification schemes can be used to automatically determine whether the uploaded video belongs to a set of copyrighted content or not.

Another application of content identification is in automatically identifying and associating content with metadata such as artist and album information. Services such as MusicBrainz [3] allow users to automatically tag their entire music collection by identifying the content. In some applications, the input may be low quality recordings (Verizon's VCAST service [4]) or short clips of the original content.

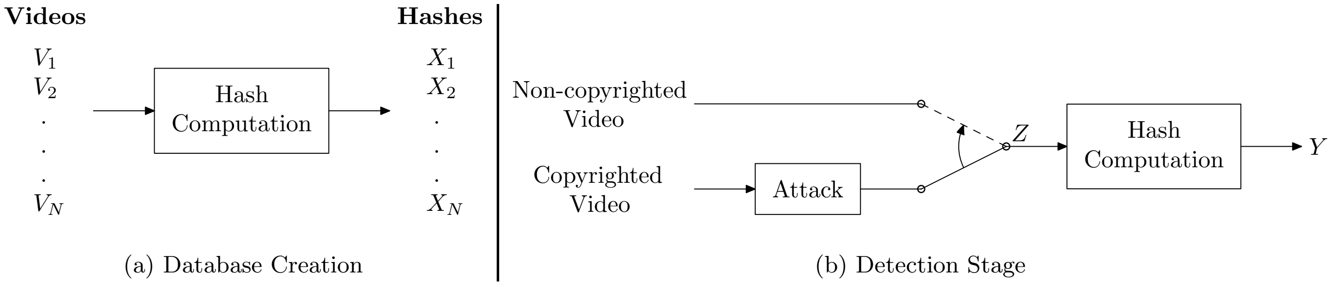
These applications demand content identification techniques that can recognize modified versions of the content, and are scalable to very large databases, often containing millions of video or audio. Existing techniques that address this problem fall into two main categories – watermarking-based techniques and hash or fingerprint-based techniques.

In the first category of watermarking-based techniques, a watermark is embedded in the host signal at the time of content creation, which can later be extracted and used to determine whether the host content is copyrighted or not, and also possibly retrieve associated metadata [6]. Watermarking has been an active area of research and robust watermarking techniques are being developed which could potentially be used for content identification. A significant limitation of watermarking-based identification systems is that a large volume of legacy multimedia content does not have any embedded watermarks and cannot be identified by watermarking-based techniques.

Hash-based techniques employ multimedia hashes designed to produce hashes that are "similar" for perceptually similar input, as opposed to cryptographic hashes which are designed to produce independent outputs if the inputs differ by even a single bit. Thus, multimedia hashes can be used to identify similar multimedia content. A hash of the uploaded content, also referred to as a content fingerprint, is computed and compared with a database of fingerprints to identify the content. Hash-based systems have the advantage that they can be used to identify existing content that does not have any embedded information. For this reason, we focus on hash-based schemes in this paper.

Several hash-based techniques have been proposed in the literature for content identification. In [11], the signs of the differences between Fourier Transform coefficients corresponding to adjacent frequency bands were used to identify audio. A similar technique for video identification was proposed in [18]. The signs of significant wavelet coefficients of spectrograms were utilized for audio identification

DRM'08, October 27, 2008, Alexandria, Virginia, USA.



(a) Database Creation

(b) Detection Stage

Figure 1: System Model

in [5]. The system in [5], called Waveprint, also uses the MinHash technique [8] combined with Locality Sensitive Hashing (LSH) [10] for fast comparison of fingerprints. Most of the existing techniques compute binary fingerprints of the content, which can be used to perform fast search over large databases.

Identification techniques proposed in the literature are currently evaluated mainly through experiments. A theoretical framework for analyzing identification schemes is necessary, as such a framework can guide the design of "optimal" identification systems. Content identification schemes employing watermarks were analyzed under a theoretical framework in [12], but the focus of the prior work was on deriving suitable statistical models for images, which can then be used to derive efficient watermark detectors. Qualitative guidelines for designing multimedia hash functions were provided in [16] by considering hashing as a source coding problem, but no quantitative analysis was provided. A decision-theoretic framework for authentication based on robust image hashing was described in [20], but is not directly applicable to the problem of content identification. In this paper, we address the content identification problem and focus on providing an analysis of binary hash-based content identification schemes. We also provide quantitative guidelines for choosing system parameters, such as the length of the hash, to achieve desired performance.

The rest of the paper is organized as follows. Our proposed framework for analyzing hash-based content identification schemes is described in Section 2. Under this framework, we examine binary hash-based content identification schemes in Section 3. We examine the applicability of the theoretical results to a practical identification scheme in Section 4. We summarize the findings and conclude in Section 5.

2. SYSTEM MODEL

The system model for a hash-based content identification scheme is shown in Fig. 1. Our system model is applicable to any hash-based identification scheme, but for ease of presentation, we illustrate our approach using the example of a video sharing website. Suppose the detector has a collection of N copyrighted videos $V_1, V_2, \ldots, V_N$. It first computes the hash $X_1, X_2, \ldots, X_N$ for each of these videos and stores them in its database. These hashes may be binary valued, such as the hashes computed in [18] and [5] or integer valued, such as color histograms or quantized SIFT features [7].

When a user uploads a video Z, the detector computes the hash Y of the uploaded video, and uses Y to decide whether the uploaded video is copyrighted or not. To evade detection, users may modify the video before uploading it to the website. Such modifications are represented by the attack block in Fig. 1(b). In these cases, the hash Y of the attacked content, may be different from the hash of the original video.

We consider two different detection scenarios based on the requirements of different applications. In some applications, such as a video sharing website implementing content filtering, it may be sufficient to determine if the content is subject to a copyright or not. In this case, the detector is only interested in determining whether a given video is present in a database of copyrighted material or not. We refer to this scenario as the *copyright detection problem*,which can be formulated as a binary hypothesis test:

$$\begin{aligned} H_0 &: Z \text{ does not correspond to a copyrighted video,} \\ H_1 &: Z \text{ corresponds to a copyrighted video.} \qquad (1) \end{aligned}$$

In some applications, such as automatic tagging of content, the detector is further interested in identifying the original video corresponding to a query video. We refer to this scenario as the *identification problem*. The identification problem can be modelled as a multiple hypothesis test with each hypothesis corresponding to one original content and a null hypothesis corresponding to the case that the uploaded video is not present in the database:

$$\begin{aligned} H_0 &: Z \text{ is not from a database } \{V_1, V_2, \ldots, V_N\}, \\ H_1 &: Z \text{ corresponds to video } V_1, \\ &\vdots \\ H_N &: Z \text{ corresponds to video } V_N. \qquad (2) \end{aligned}$$

In the next section, we analyze binary hash-based schemes under this framework and derive a lower bound on the length of the hash required to achieve low error probabilities.

3. ANALYSIS OF BINARY HASH SCHEMES

Binary strings are commonly employed in hashing schemes, as comparison of binary strings can be performed efficiently. In this section, we examine both the detection and identification problems for binary hash-based content identification schemes.

3.1 Hash and Attack Models

From the designer's point of view, it is desirable for the hash bits to be independent of each other, so that an attacker cannot alter a significant number of hash bits at once by making minor changes to the content. Further, if the hash bits are equally likely to be 0 or 1, the overall entropy is maximized and each bit conveys the maximum amount of information. If the hash bits are not equally likely to be 0 or 1, they could be compressed into a shorter vector with equiprobable bits. Therefore, in our analysis, we assume that the bits comprising the hash are independent of each other and are 0 or 1 with probability 0.5.

A user wishing to upload copyrighted content onto a website, may make modifications to the content so as to evade detection. These attacks on the image content are reflected as changes in the computed fingerprint. By a suitable choice of hash features and appropriate preprocessing and synchronization, such attacks can be modeled as additive noise in the hash space [16]. We illustrate the importance of proper choice of hash features in Section 4.5. For the remainder of the paper, we assume that the attacks can be represented as additive noise applied to the original hash.

Since the hash bits are designed to be independent and identically distributed (i.i.d.), we model the effect of attacks on the multimedia content as altering each bit of the hash independently with probability $p < 0.5$. The maximum possible value of p is linked to the maximum amount of distortion an attacker may introduce into the multimedia content.

3.2 Detection Problem

Under the assumptions outlined above, the *copyright detection problem* becomes:

$$\begin{aligned} H_0 &: Y \notin \{X_1, X_2, \ldots, X_N\} + n, \\ H_1 &: Y = X_i + n, \text{ for some } i \in \{1, 2, \ldots, N\}. \end{aligned} \tag{3}$$

where Y, $X_i, i = 1, 2, \ldots, N$ and the noise n are all binary vectors of length L. Under hypothesis H_0, Y can take any value with equal probability, since the hash bits are designed to be i.i.d. with equal probability of being 0 or 1. Thus, under H_0, $p(Y|H_0) = \frac{1}{2^L}$. The distribution of the hash Y, given that it is a modified version of hash X_i, $p(Y|X_i)$ can be specified by considering its Hamming distance. Let $d_i = d(Y, X_i)$ be the Hamming distance between the hash of the video being uploaded and a hash X_i in the database. Since the probability of a bit flipping is p, the probability that exactly d_i bits are altered is $p(Y|X_i) = p^{d_i}(1-p)^{L-d_i}$.

The alternative hypothesis, H_1, is thus a composite hypothesis, as the computed hash can have different distributions depending on which original hash it corresponds to. The optimal decision rule for composite hypothesis testing is given as [19]:

$$\text{Decide } H_1 \text{ if } \quad \frac{p(Y|H_1)}{p(Y|H_0)} > \tau'', \tag{4}$$

where the threshold τ'' can be chosen to satisfy some optimality criterion. If the priors of the hypotheses and the associated costs are known, then τ'' can be computed so as to minimize the expected Bayes' risk. If the costs are known, but not the priors, the threshold τ'' can be chosen to minimize the maximum expected risk. In this paper, we use a Neyman-Pearson approach [19] to maximize the probability of detection P_d subject to the constraint that the probability of false alarm $P_f \leq \alpha$.

To simplify the analysis, we assume that all copyrighted videos are equally likely to be uploaded. In situations where some videos may be uploaded more often than those which are less popular, the analysis can be extended by appropriately modifying the prior probabilities. With this assumption, the likelihood ratio test in Eqn. (4) becomes:

$$\frac{\sum_{i=1}^{N} p(Y|X_i)p(X_i|H_1)}{p(Y|H_0)} > \tau'',$$

Substituting $p(Y|H_0) = \frac{1}{2^L}$, $p(Y|X_i) = p^{d_i}(1-p)^{L-d_i}$, and $p(X_i|H_1) = \frac{1}{N}$, we get:

$$\sum_{i=1}^{N} \left(p^{\frac{d_i}{L}} (1-p)^{1-\frac{d_i}{L}} \right)^L > \tau'. \tag{5}$$

where the constants have been absorbed into the threshold τ'. We note that the left hand side is a sum of exponentials, and for large L, only the largest term would be relevant. Further, since $p^x(1-p)^{1-x}$ is a decreasing function of x for $p < 0.5$, the largest term in the left hand side of Eqn. (5) would be the one with the smallest value of d_i. Thus, we arrive at the decision rule:

$$\begin{cases} \text{Decide } H_1 & \text{if} \quad d_{\text{min}} < \tau, \\ \text{Decide } H_1 \text{ with probability } q & \text{if} \quad d_{\text{min}} = \tau, \\ \text{Decide } H_0 & \text{if} \quad d_{\text{min}} > \tau, \end{cases} \tag{6}$$

where $d_{\text{min}} = \min_{i=1,2,\ldots,N} d_i$ and τ an integer threshold expressed in terms of the Hamming distance. q and τ are chosen to achieve a desired probability of false alarm α. Based on this decision rule, we decide that the uploaded video is a modified version of some video in the database if the distance of the hash of the uploaded video to the closest hash in the database is less than a threshold.

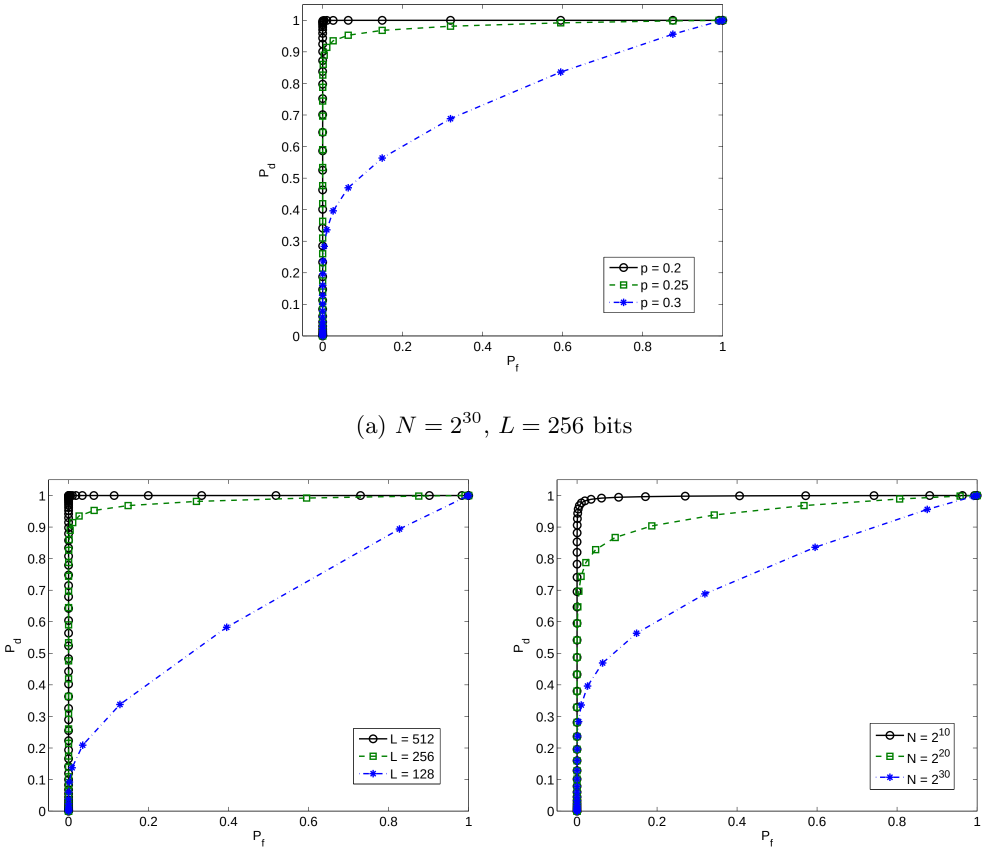
The corresponding probability of false alarm P_f for a threshold τ is given by $P_f = \Pr(d_{\text{min}} < \tau|H_0) + q\Pr(d_{\text{min}} = \tau|H_0)$, and the probability of detection is given as $P_d = \Pr(d_{\text{min}} < \tau|H_1) + q\Pr(d_{\text{min}} = \tau|H_1)$. The final expressions for P_d and P_f are derived in Appendix A.

In Fig. 2, we show the receiver operating characteristics (ROC) [19] computed using the expressions derived in Appendix A, for various values of the parameters L, N, and p. Fig. 2(a) shows the ROC curves as the strength of the attack p is increased from 0.2 to 0.3 for $N = 2^{30}$ videos in the database and a hash length of 256 bits. We observe that as the attack strength p increases, the probability of detecting a copyrighted video, P_d, reduces, for a given probability of false alarm P_f. As p approaches 0.5, the probability of detection approaches the lower bound $P_d = P_f$. Fig. 2(b) shows that for a given attack strength, the detector performance can be improved by using a longer hash. As the hash length is increased, P_d increases for a given P_f. Fig. 2(c) shows the influence of the number of videos in the database, N, on the detector performance for a fixed hash length $L = 256$ bits and attack strength $p = 0.3$. As N increases, the probability of false alarm increases. Hence, for a given P_d, the P_f is higher, or equivalently, for a fixed P_f, the probability of detection is lower.

3.3 Identification Problem

We now consider the *identification problem* for binary hash-based schemes, which can be modelled as a multiple hypothesis test:

$$\begin{aligned} H_0 &: Y \notin \{X_1, X_2, \ldots, X_N\} + n, \\ H_1 &: Y = X_1 + n, \\ &\vdots \\ H_N &: Y = X_N + n. \end{aligned} \tag{7}$$



(a) $N = 2^{30}$, $L = 256$ bits

(b) $N = 2^{30}$, $p = 0.3$

(c) $p = 0.3$, $L = 256$ bits

Figure 2: Receiver Operating Characteristics (ROC) for the binary hypothesis testing problem obtained from theoretical analysis.

We define the probability of correct identification, P_c, as the probability of correctly identifying the original video corresponding to an uploaded video. As the hashes $X_i, i = 1, 2, \ldots, N$ are identically distributed, and the distribution of the noise n under each of the hypotheses is the same, the overall probability of correct identification P_c will be equal to the probability of correct identification under a given hypothesis i, $P_c = \Pr(\text{deciding } H_i | H_i \text{ is true})$, $i \neq 0$. The probability of misclassification, P_m, can be obtained as $P_m = \Pr(\text{deciding } H_i, i \neq j \neq 0 | H_j \text{ is true})$. The probability of falsely classifying a non-copyrighted video as copyrighted (false alarm) is defined as $P_f = \Pr(\text{deciding } H_i | H_0 \text{ is true})$, $i \neq 0$.

As before, we assume that the hash bits are i.i.d. and equally likely to be 0 or 1. The noise is modelled as independently altering each bit with probability p. Under this model, the Maximum Likelihood (ML) decision rule can be derived as:

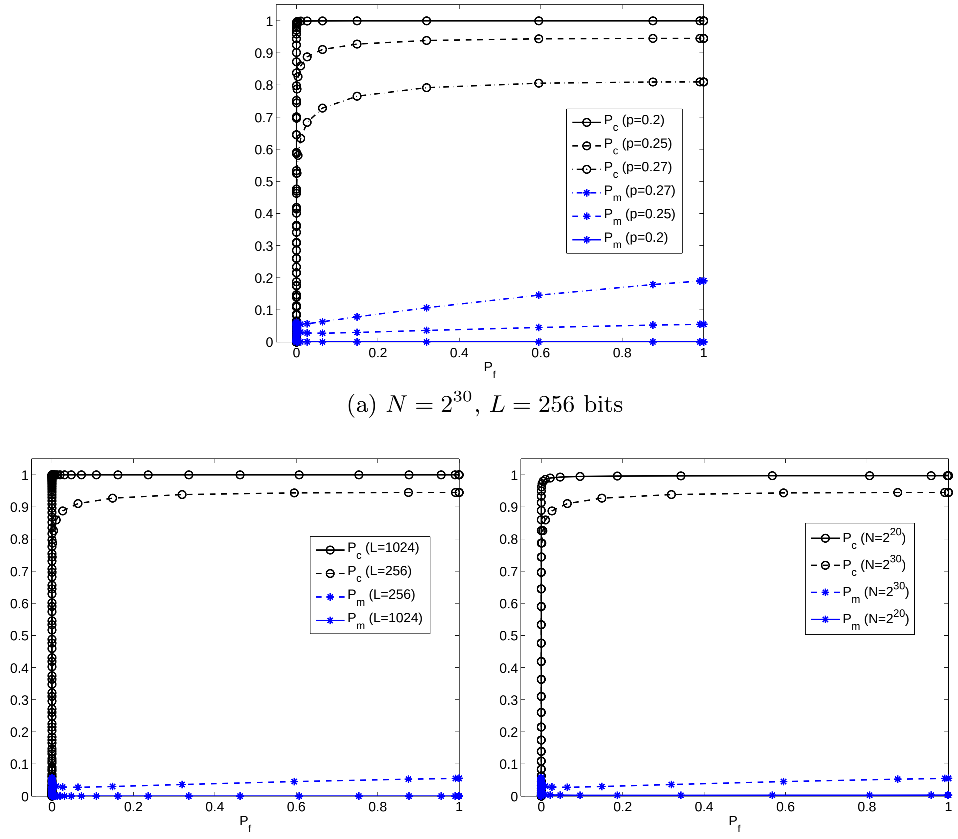
$$\begin{cases} \text{decide } H_i & \text{if } \left(i = \underset{j=1,2,\ldots,N}{\arg\min}\, d_j \right) \text{ and } d_i \leq \tau, \\ \text{decide } H_0 & \text{otherwise.} \end{cases} \qquad (8)$$

If hashes of several copyrighted videos have the same distance to the hash of the uploaded video Y, one of them is chosen randomly. For this ML detector, the expressions for the performance metrics P_c, P_m, and P_f can be derived as shown in Appendix B.

Fig. 3 shows the influence of the various parameters on the performance metrics for the ML detector in Eqn. (8). Fig. 3(a) shows the influence of the attack strength p. We observe that as p increases, the probability of correct identification P_c at a given false alarm probability P_f reduces, and the probability of misclassification P_m increases. Fig. 3(b) shows that the probability of correct identification under a given attack strength p and a given P_f can be increased by increasing the hash length. The influence of the number of videos N on the accuracy of identification is shown in Fig. 3(c). As the number of videos in the database increases, the probability of false alarm increases, or equivalently, at a given P_f, the value of P_c is lower. These results are similar to that obtained for the detection problem (Eqn. (3)). Thus, given the number of videos N and a desired probability of false alarm P_f, the content identification system can be made more robust by choosing a longer hash length L.

3.4 Bound on the Required Hash Length

As shown in the previous sections, the length of the hash is an important parameter that controls the performance of the system. We now derive a lower bound on the length of the hash, as a guideline for choosing the length to achieve



(a) $N = 2^{30}$, $L = 256$ bits

(b) $N = 2^{30}$, $p = 0.25$

(c) $p = 0.25$, $L = 256$ bits

Figure 3: ROC curves for the multiple hypothesis testing problem obtained from theoretical analysis.

a desired probability of false alarm $P_f = \epsilon$ given a certain number of videos N, with probability of detection $P_d \to 1$. We present results for the detection problem and similar results can be derived for the identification problem.

For simplicity, we set $q = 1$ in Eqn (6). As derived in Appendix A (Eqn. (14)),

$$\epsilon \quad = \quad 1 - \left(1 - \frac{1}{2^L} \sum_{j=0}^{\tau} \binom{L}{j}\right)^N.$$

For small ϵ,

$$\frac{\epsilon}{N} \approx \frac{1}{2^L} \sum_{j=0}^{\tau} \binom{L}{j}.$$

For large L, it can be shown that

$$\frac{1}{L+1} 2^{Lh(\lambda)} \leq \sum_{j=0}^{L\lambda} \binom{L}{j} \leq 2^{Lh(\lambda)}.$$

Using the above result, we obtain

$$\frac{1}{L} \log_2 \frac{N}{\epsilon} \approx 1 - h\left(\frac{\tau}{L}\right), \qquad (9)$$

where $h(x)$ is the binary entropy function given by $h(x) = -x \log_2 x - (1-x) \log_2 (1-x)$.

Let $\Phi(k) = \sum_{j=0}^{k} \binom{L}{j} p^j (1-p)^{L-j}$ be the cumulative distribution function (c.d.f.) of a binomial random variable with parameters L and p. It can be shown that

$$\begin{aligned} P_d \quad &\approx \quad 1 - \left[\left\{1 - \frac{(N-1)\epsilon}{N}\right\}(1 - \Phi(\tau))\right], \\ &= \quad \Phi(\tau) + \frac{(N-1)\epsilon}{N}(1 - \Phi(\tau)). \qquad (10) \end{aligned}$$

Since ϵ is small, to have $P_d \to 1$, we require $\Phi(\tau) \to 1$, for which a necessary condition is $\tau > Lp$. Combining this condition with Eqn. (9), we obtain the following important result:

THEOREM 1. *Given a database of N videos and an attack strength $p < \frac{1}{2}$, a desired probability of false alarm $P_f = \epsilon$ can be achieved with the probability of detection $P_d \approx 1$, by choosing a large enough hash length L that satisfies the bound*

$$L > \frac{1}{1 - h(p)} \log_2 \frac{N}{\epsilon}, \qquad (11)$$

or equivalently,

$$\frac{1}{L} \log_2 \frac{N}{\epsilon} < 1 - h(p). \qquad (12)$$

Fig. 4 shows the variation of the lower bound on the length of the hash required to resist an attack strength p under different requirements on the probability of false alarm P_f. We observe that with a database of approximately one billion (2^{30}) videos and a desired false alarm probability of $2^{-50} \approx 10^{-15}$, a hash length of approximately 3000 bits suffices to resist strong attacks that alter 40% of the hash bits.

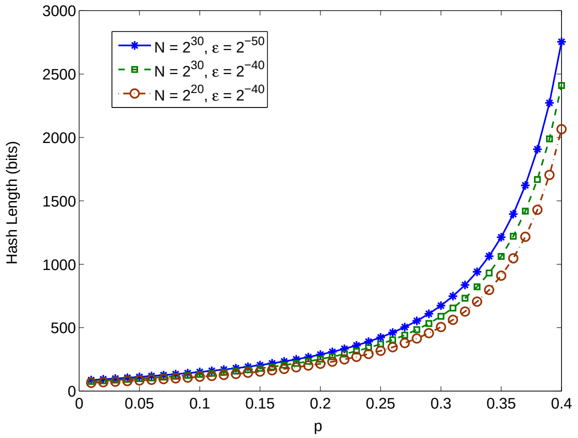


Figure 4: Lower bound on the length of the hash required to resist an attack which alters hash bits with probability p, under different settings.

4. EVALUATION OF A PRACTICAL HASH SCHEME

In this section, we examine the applicability of our theoretical results to a practical identification scheme. We illustrate using a simple image hashing scheme based on the wavelet transform coefficients [15]. A similar scheme for video hashing based on DCT coefficients has been proposed in [9]. We present results for image identification, but the results can be easily extended to the case of video or audio identification using schemes such as [9].

4.1 Hash Generation

Wavelet coefficients, and in particular, signs of wavelet coefficients have been used for content identification [5], retrieval of similar images [13], and to generate hashes for image authentication [17]. It has been shown that detail coefficients of the wavelet transform are symmetric around zero and can be modelled as i.i.d. generalized Gaussian random variables [14]. Thus, quantizing wavelet detail coefficients to 1 bit would yield i.i.d. equiprobable bits, which could be used as fingerprints to represent the image.

We decompose a 512×512 image up to five levels using the Haar wavelet [15], which is chosen because of the low cost for computing the transform. Each of the four subbands at the coarsest level of decomposition thus has coefficients of size 16×16. We retain only the signs of the coefficients belonging to these subbands to obtain a 1024-bit sequence. A '1' at a particular location indicates a positive coefficient, whereas a '0' indicates a negative coefficient. Fig. 5 shows the distribution of the bits comprising this bit sequence estimated from 1000 grayscale images of size 512×512. In Fig. 5(a), we show the fraction of images (out of 1000) that have a '1' at a particular location. The first 256 bits correspond to the signs of the approximation coefficients, followed by 256 bits for each of the horizontal, vertical and diagonal detail coefficients. From this figure, we observe that the signs of the approximation coefficients are not independent and equally likely. This is due to the fact that the approximation coefficients for natural images are likely to be correlated with each other. The same holds true for the horizontal and vertical detail coefficients, since coefficients which correspond to strong horizontal or vertical edges would lie along the same row or column, respectively. The signs of the diagonal detail coefficients, however, appear to be less correlated and approximately equally likely to be '0' or '1'. Fig. 5(b) shows the fraction of bits that are '1' for a given image. We observe that approximately half the bits are '1', indicating that these bits are approximately independent and equally likely. The coefficients at the lowest level of decomposition are also expected to be robust to common signal processing operations and can be used as hashes for image identification.

Thus, given an image, we resample it to size 512×512, perform wavelet decomposition up to 5 levels, and extract the diagonal detail coefficients. We then retain the signs of these coefficients to form a 256-bit hash for the given image.

4.2 Attacks

We evaluate the ability of these hashes to correctly identify an image after it has undergone the potential malicious attacks listed in Table 1. As the image pixel values are normalized to lie between 0 and 1, addition of zero mean Gaussian noise with standard deviation $\sigma = 0.2$ represents a strong attack and introduces a lot of distortion, as shown in Fig. 6. Rotation by multiples of 90° (Attack No. 32-34) are very strong attacks that may be of concern if the image/video is being viewed on a portable device, which provides freedom in adjusting the orientation.

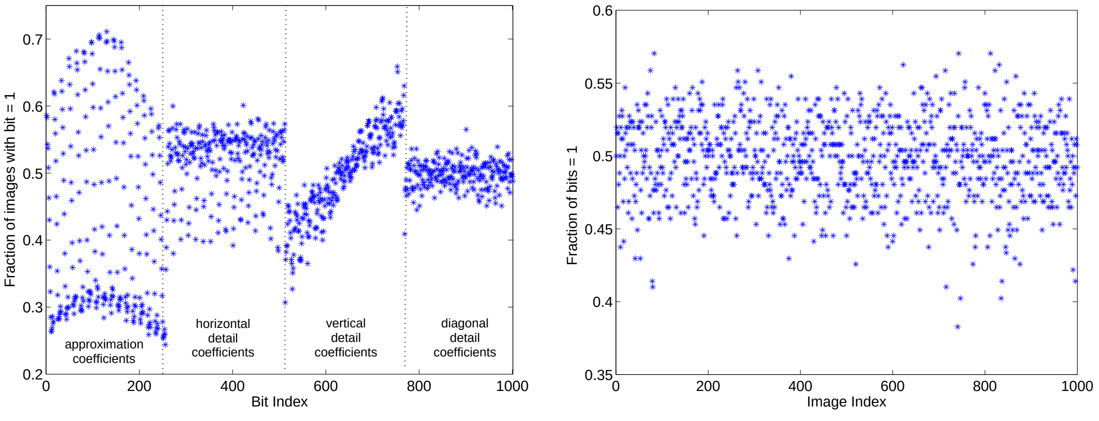
The strength of an attack can be measured in terms of the probability (p) of a hash bit being altered after the attack. Fig. 7 shows the probability of a hash bit being changed as a result of each attack, averaged over 1000 images. We observe that the rotation attacks are devastating, and the probability of a hash bit being altered is almost 0.5 for each of them. Theorem 1 suggests that the hashing scheme will not accurately identify the images after these attacks due to the high value of p. Among the other attacks, Gaussian noise addition with standard deviation of 0.2 (Attack No. 4) causes the highest number of changes to the hash bits.

4.3 Performance Evaluation

We now evaluate the accuracy of the content identification system under these attacks. Our database consists of $N = 1000$ grayscale images of size 512×512. The attacks in Table 1 are applied to each of these images to obtain a set of $34,000$ attacked images. The length of the hash used is $L = 256$ bits. The threshold for detection τ is chosen to achieve a probability of false alarm $\epsilon = 10^{-6}$. From Eqn. (11), the maximum attack strength that can be resisted under these settings is found to be $p = 0.3$. Thus, we expect that the rotated images (Attack No. 32-34) which have $p = 0.5$ will not be detected correctly. The other attacks no. $1 - 31$ have $p < 0.3$ and hence we expect the probability of detection P_d to be close to 1.

For the detection problem, we compute the hash of an attacked image and compare it with each hash in the database. We then use the decision rule described in Eqn. (6) to perform the classification. If the minimum distance $d_{\min} < \tau$, we declare the image to be present in the database. Fig. 8 shows the probability of detection obtained using this decision rule under each of the attacks. As expected, the images which correspond to rotated versions of images in the database are almost never detected (Attacks No. $32 - 34$). This problem can be alleviated by suitably designing the hashes, as discussed in Section 4.5.

Under most of the other attacks, the probability of detection P_d is close to 1, except for addition of Gaussian noise with large variance (Attacks No. 2-4). Under these attacks,

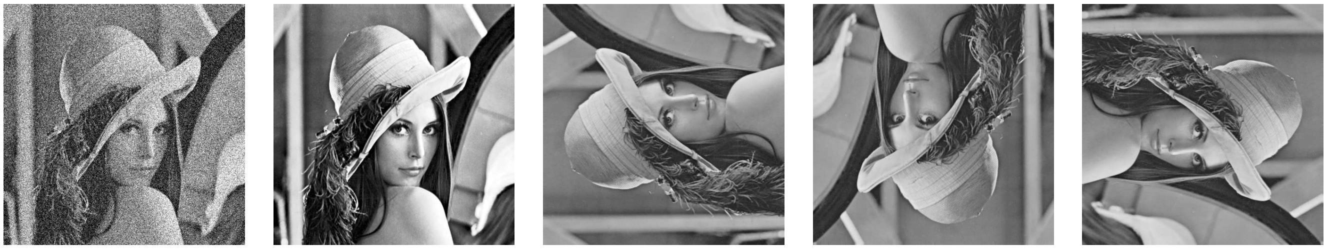


(a) (b)

Figure 5: Distribution of hash bits obtained by quantizing wavelet coefficients. (a) Fraction of images with a bit '1' at a given location and (b) Fraction of bits that are '1' for a given image.

Table 1: List of attacks tested.

Attack No.	Attack	Parameters
1-4	Zero-mean Gaussian Noise Addition	$\sigma = 0.05, 0.1, 0.15, 0.2$
5-8	Uniform Noise Addition $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$	$\Delta = 0.05, 0.1, 0.15, 0.2$
9	Histogram Equalization	
10-19	Gamma Correction	$\gamma = 0.75 : 0.05 : 1.25$ 1
20-28	Average, Median, and Gaussian Filtering	Filter Size $= 3, 5, 7$
29-31	JPEG Compression	Quality Factor $= 25, 50, 75$
32-34	Rotation by multiples of 90°	



Attack 4 Attack 9 Attack 32 Attack 33 Attack 34

Figure 6: Some attacked versions of the Lena image. The list of attacks is provided in Table 1.

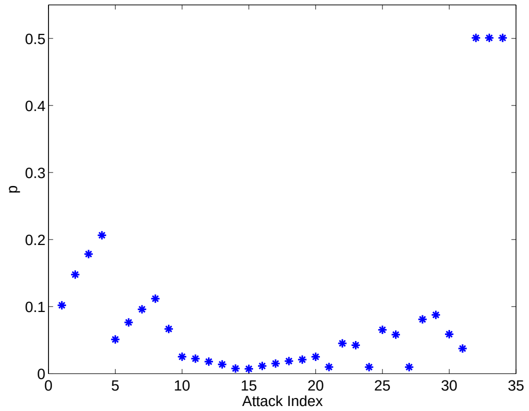


Figure 7: Probability of a hash bit flipping averaged over 1000 images for each attack.

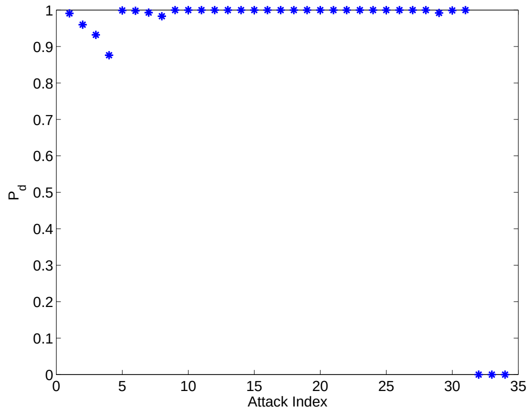


Figure 8: Probability that an attacked image is detected as a modified version of an image in the database P_d.

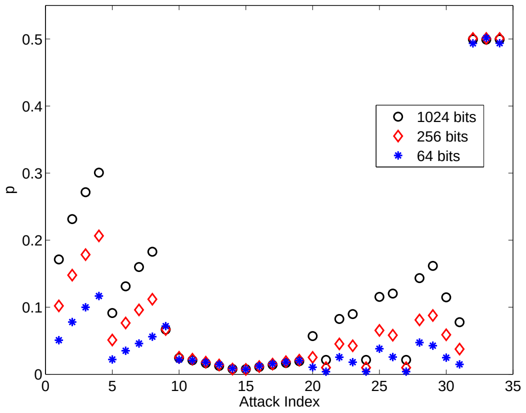


Figure 9: Probability of a hash bit flipping under an attack on the image as a function of the hash length.

the fraction of hash bits altered for some images is larger than 0.3. Thus, according to our theoretical analysis, these images cannot be identified and the probability of detection, P_d, is less than 1 for these attacks. The overall probability of detection for attacks no. 1-31 was 0.991.

For the identification problem, we use the ML detector in Eqn. (8) to perform the classification. We found that *every* image that was detected as being present in the database in the detection problem was correctly identified, so that $P_c = 0.991$ and the probability of misclassification $P_m = 0$.

The probability of false alarm P_f was estimated using the leave-one-out procedure in both the detection and identification problems. Every image in the database was treated as a probe image and compared with the remaining images. If the minimum distance of the hash $d_{\min} < \tau$, the image constituted a false alarm. Using the hash of length 256 bits, no false alarms were observed in our experiments.

4.4 Influence of the Hash Length

From Theorem 1, we know that longer hashes can resist stronger attacks. In this subsection, we perform simulations to determine the influence of the hash length on the detection performance.

To generate hashes of different lengths, the number of levels of the wavelet decomposition is varied. For example, to generate hashes of length 1024 bits, we resample the image to size 512 × 512 and decompose it to four levels using the Haar wavelet. We then extract the signs of the diagonal detail coefficients at the coarsest level of decomposition. As the number of levels of decompositions becomes smaller, the diagonal detail coefficients correspond to higher frequencies and we expect these features to be less robust to modifications. Fig. 9 shows the probability of a hash bit flipping after attacks, for hashes of length 64, 256, and 1024 corresponding to 6, 5, and 4 levels of decomposition respectively. We observe that as the number of decomposition levels decreases (corresponding to longer hashes), the probability that a hash bit changes increases, indicating that these coefficients are less robust to modifications.

From our theoretical analysis (Eqn. (12)), we find that for $N = 1000$ and $\epsilon = 10^{-6}$, the maximum probability of a bit flipping (p) that can be tolerated by hashes of length 64, 256, and 1024 bits is 0.1, 0.3, and 0.4, respectively. Thus, we expect the hash with length 1024 bits to have a higher value of P_d, as it can resist stronger attacks.

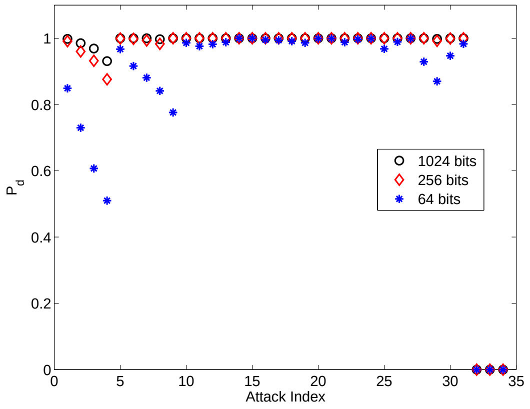


Figure 10: Probability of detection under various attacks as a function of the hash length.

Table 2: Overall P_d and P_f obtained against Attacks No. 1-31 as a function of the hash length.

Hash Length (bits)	P_d	P_f
64	0.924	0
256	0.991	0
1024	0.996	0.002

In Fig. 10, we examine the influence of the hash length L on the probability of detection P_d under various attacks. In each case, the threshold for detection τ was chosen to attain the desired value of $P_f = \epsilon = 10^{-6}$ as given by Eqn. (9). We observe that the hash with length 1024 bits has the highest probability of detection. Even though the probability of a bit being altered after an attack p is higher for the 1024-bit hash than the other hashes, the longer length of the hash compensates for the reduced robustness of each individual bit, and leads to a higher overall probability of detection.

In Table 2, we compare the overall probability of detection under attacks no. $1-31$ as a function of the hash length. We observe that as the hash length increases, P_d also increases. There was only one case of false alarm when using hashes of length 1024 bits. Upon closer observation, it was found that these two images actually corresponded to the same scene, but the number of objects and illumination conditions in the picture were slightly different. These two images can be regarded as being obtained from each other after significant modification, such as insertion or deletion of objects, change in brightness, and modification of the details in the image. The overall attack would change a large fraction of the hash bits, and is hence not identified using the shorter hashes. Since the 1024 bit hash is more robust against changes in the hash bits, it is able to determine that these two images are not independent of each other, and could have originated from the same source. Thus, the length of the hash plays a crucial factor in determining the performance of the hashing scheme, as predicted by our theoretical analysis in Section 3.

Under the identification problem, every image that is detected as having originated from an image present in the database is also correctly identified, so that $P_c = P_d$. Thus, the probability of misclassification as obtained from our experiments is $P_m = 0$ and the probability of correct identification is the same as the second column in Table 2.

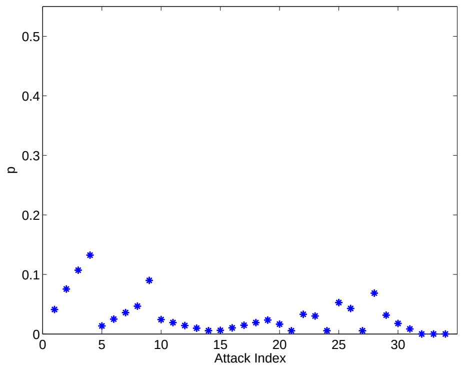


Figure 11: Probability of a hash bit flipping p for the rotationally invariant hashes under various attacks.

4.5 Proper Choice of Hash Features

Our attack model assumes that most attacks on multimedia can be modelled as additive noise in the hash space. For some hashing schemes, desynchronization attacks, including rotation, cropping, and geometric attacks, may not be directly modelled as additive noise hash space. However, by suitably designing the hash features and applying appropriate preprocessing, it is possible to reduce these attacks to the additive noise model. We briefly illustrate the importance of appropriate choice of hash features using the example of the rotation attacks studied in Section 4.2. If robustness against rotations by multiples of 90° is desired, the following modification of the hash scheme in Section 4.1 can improve the robustness against rotations.

Given a 512×512 image, we obtain four images corresponding to rotations by multiples of 90°, which are then summed pixelwise. The resulting image is decomposed up to four levels using the Haar wavelet and the signs of the 1024 diagonal detail coefficients at the coarsest level of decomposition are extracted. As these bits are dependent, we retain only 25% of the bits that correspond to the coefficients in the upper left corner of the subband. The 256 bits thus obtained form the hash for the image, which is invariant under rotations of the original image by multiples of 90°.

Fig. 11 shows the probability of a hash bit flipping under the attacks listed in Table 1 for this modified hash scheme. We observe that none of the hash bits are altered under rotations by multiples of 90°. The hash bits are also moderately robust under the other attacks no. 1-31. Under the detection problem we obtained $P_d = 1$ under the rotation attacks no. 32-34, while the overall P_d for attacks no. $1-31$ was 0.99. Thus, a suitable choice of the hash features can enhance the robustness against attacks.

5. CONCLUSIONS

In this paper, we have presented a decision theoretic framework for analyzing binary hash-based content identification schemes. We formulate the problem of detecting whether a given video or audio is present in a database of copyrighted material as a binary hypothesis test and the problem of correctly identifying the original content corresponding to a given query object as a multiple hypothesis test. Under this framework, we have derived expressions for the relevant performance metrics such as probability of detection, the probability of correct identification and the probability of false alarm. We have obtained a lower bound on the length of the hash required to resist a given attack strength.

Under the proposed framework, we also examined a practical binary hash-based content identification scheme which utilizes the signs of the diagonal detail coefficients in the wavelet decomposition of the image. The simulation results confirm our theoretical predictions. We also briefly discussed the importance of choosing appropriate hash features to achieve robustness against attacks.

As future work, we plan to extend the theoretical results derived for binary hashes to other hash-based content identification systems and use the analytic results to guide the design of hash-based schemes with higher accuracy of identification.

6. REFERENCES

[1] Youtube: http://www.youtube.com.

[2] Wall Street Journal: YouTube Removes 30,000 Files Amid Japanese Copyright Concerns. http://online.wsj.com/article/SB116133637777798831.html.

[3] MusicBrainz: http://www.musicbrainz.org/.

[4] Verizon VCAST Song ID: http://solutions.vzwshop.com/songid/.

[5] S. Baluja and M. Covell. Content Fingerprinting using Wavelets. In *Proc. of IET Conf. on Multimedia*, London, England, November 2006.

[6] M. Barni and F. Bartolini. Data Hiding for Fighting Piracy. *IEEE Signal Processing Magazine*, 21(2):28–39, March 2004.

[7] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable Near Identical Image and Shot Detection. In *ACM Int'l Conf. on Image and Video Retrieval*, Amsterdam, July 2007.

[8] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. Ullman, and C. Yang. Finding Interesting Associations without Support Pruning. *IEEE Trans. on Knowledge and Data Engineering*, 13(1):64–78, January 2001.

[9] B. Coskun, B. Sankur, and N. Memon. Spatio-temporal Transform Based Video Hashing. *IEEE Trans. on Multimedia*, 8(6):1190–1208, Dec. 2006.

[10] A. Gionis, P. Indyk, and R. Motwani. Similarity Search in High Dimensions via Hashing. In *Proc. of the Int'l Conf. on Very Large Databases*, pages 518–529, 1999.

[11] J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System. In *Proc. of the Int'l Symposium on Music Information Retrieval*, Paris, France, 2002.

[12] J. Hernandez and F. Perez-Gonzalez. Statistical Analysis of Watermarking Schemes for Copyright Protection of Images. *Proc. of the IEEE*, 87(7):1142–1166, July 1999.

[13] C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast Multiresolution Image Querying. In *Proc. of the 22nd Annual Conf. on Computer Graphics and Interactive Techniques*, pages 277–286, New York, USA, 1995.

[14] S. Mallat. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[15] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, second edition, 1999.

[16] E. McCarthy, F. Balado, G. Slvestre, and N. Hurley. A framework for Soft Hashing and its Application to Robust Image Hashing. In *IEEE Int'l Conf. on Image Proc.*, volume 1, pages 397–400, Oct. 2004.

[17] M. K. Mihçak and R. Venkatesan. New Iterative Geometric Methods for Robust Perceptual Image Hashing. In *ACM Workshop on Security and Privacy in Digital Rights Management*, 2001.

[18] J. Oostveen, T. Kalker, and J. Haitsma. Feature Extraction and a Database Strategy for Video Fingerprinting. In *Proc. of the 5th Int'l Conf. on Recent Advances in Visual Information Systems, Lecture Notes in Computer Science*, volume 2314, pages 117–128, 2002.

[19] H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer, second edition, 1994.

[20] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun. Robust Perceptual Hashing as Classification Problem: Decision-theoretic and Practical Considerations. In *IEEE Workshop on Multimedia Signal Processing*, pages 345–348, Oct. 2007.

APPENDIX

A. PROBABILITIES OF ERROR FOR DETECTION PROBLEM

As discussed in Section 3, the copyright detection problem can be formulated as a binary hypothesis test. We now derive expressions for the probability of false alarm, P_f, and the probability of detection, P_d, under this setting.

The probability of false alarm, P_f, is given by

$$\begin{aligned} P_f &= \Pr(d_{\min} < \tau | H_0) + q \Pr(d_{\min} = \tau | H_0), \\ P_f &= 1 - [\Pr(d_i \geq \tau | H_0)]^N + \\ & q \sum_{j=1}^{N} \binom{N}{j} [\Pr(d_i = \tau | H_0)]^j [\Pr(d_i > \tau | H_0)]^{N-j}. \quad (13) \end{aligned}$$

The hamming distance between Y and X_i $d_i = d(Y, X_i) = \text{wt}(Y + X_i)$, where $\text{wt}(\cdot)$ denotes the Hamming weight of a binary vector and + denotes modular addition over the binary field. Under H_0, since Y is uniformly distributed over the entire space of binary strings of length L, $Y + X_i$ is also uniformly distributed. The probability distribution of $\text{wt}(Y + X_i)$ thus corresponds to the weight of a uniformly distributed random vector, which is a binomial distribution with parameters L and 0.5. Thus,

$$\begin{aligned} \Pr(d_i = \tau | H_0) &= \frac{1}{2^L} \binom{L}{\tau}, \\ \Pr(d_i < \tau | H_0) &= \frac{1}{2^L} \sum_{j=0}^{\tau-1} \binom{L}{j}. \end{aligned}$$

Substituting the above expressions into Eqn. (13), the probability of false alarm can be written as

$$P_f = 1 - \left[\frac{1}{2^L} \sum_{j=\tau}^{L} \binom{L}{j} \right]^N +$$

$$\frac{q}{2^{LN}} \times \sum_{j=1}^{N} \binom{N}{j} \left[\binom{L}{\tau} \right]^j \left[\sum_{k=\tau+1}^{L} \binom{L}{k} \right]^{N-j}.$$

To simplify the notation, define $f_0(k) = \frac{1}{2^L} \binom{L}{k}$ and $f_1(k) = \binom{L}{k} p^k (1-p)^{L-k}$. Also define $F_0(k) = \sum_{j=k}^{L} f_0(j)$ to be the tail probability of the binomial distribution with parameters L and $\frac{1}{2}$, and $F_1(k) = \sum_{j=k}^{L} f_1(j)$. The probability of false alarm can now be written as:

$$\begin{aligned} P_f &= 1 - [F_0(\tau)]^N + \\ & q \sum_{j=1}^{N} \binom{N}{j} [f_0(\tau)]^j [F_0(\tau+1)]^{N-j}, \\ &= 1 - (1-q)[F_0(\tau)]^N - q[F_0(\tau+1)]^N. \quad (14) \end{aligned}$$

The probability of detection is given as

$$P_d = \Pr(d_{\min} < \tau | H_1) + q \Pr(d_{\min} = \tau | H_1).$$

Suppose H_1 is true and that the uploaded video is actually an attacked version of copyrighted video V_s. Then we have, $\Pr(d_s = d) = \binom{L}{d} p^d (1-p)^{L-d}$. For $i \neq s$, since the X_is are uniformly distributed over the entire space, $Y + X_i$ is also uniformly distributed. The distance $d_i = \text{wt}(Y + X_i)$ follows a binomial distribution with parameters L and 0.5. Thus, $\Pr(d_i = d, i \neq s) = \frac{1}{2^L} \binom{L}{d}$ and the probability of detection can be derived as

$$\begin{aligned} P_d &= 1 - [F_1(\tau)][F_0(\tau)]^{N-1} + q f_1(\tau) [F_0(\tau)]^{N-1} + \\ & q[F_1(\tau+1)] \sum_{j=1}^{N-1} \binom{N-1}{j} [f_0(\tau)]^j [F_0(\tau+1)]^{N-1-j}, \\ &= 1 - (1-q)[F_1(\tau)][F_0(\tau)]^{N-1} - q[F_1(\tau+1)][F_0(\tau+1)]^{N-1}. \end{aligned}$$

B. PROBABILITIES OF ERROR FOR IDENTIFICATION PROBLEM

We now compute the performance metrics for the ML detector (Eqn. (8)) under the identification problem (Eqn.(2)). The probability of false alarm is given by

$$\begin{aligned} P_f &= \Pr(\text{at least one of } d_1, d_2, \ldots, d_N \leq \tau | H_0), \\ &= 1 - \Pr(\text{none of } d_1, d_2, \ldots, d_N \leq \tau | H_0), \\ &= 1 - [F_0(\tau+1)]^N. \end{aligned}$$

As the distribution of the noise and the hashes under the different hypotheses are identical, the overall probability of correct classification is equal to the probability of correctly identifying video V_s given as:

$$\begin{aligned} P_c &= \Pr(\text{deciding } s | H_s) \\ &= \Pr(d_s \leq \tau \text{ and } d_s < \min_{i \neq s} d_i | H_s) + \\ & \Pr(\min_{i \neq s} d_i = d_s \leq \tau \text{ and } s \text{ is decided} | H_s), \\ &= \sum_{j=0}^{\tau} f_1(j) \Big[\{F_0(j+1)\}^{N-1} + \\ & \sum_{k=1}^{N-1} \frac{1}{k+1} \binom{N-1}{k} [f_0(j)]^k [F_0(j+1)]^{N-1-k} \Big]. \end{aligned}$$

Similarly, the probability of misclassification can be computed as:

$$\begin{aligned} P_m &= \Pr(\text{deciding } i \neq s \neq 0 | H_s), \\ &= \Pr(\min_{i \neq s} d_i \leq \tau \text{ and } \min_{i \neq s} d_i < d_s | H_s) + \\ & \Pr(\min_{i \neq s} d_i = d_s \leq \tau \text{ and } i \neq s \text{ is decided} | H_s), \\ &= \sum_{j=0}^{\tau} \Bigg[\sum_{k=1}^{N-1} \Bigg\{ \binom{N-1}{k} f_0(j)^k [F_0(j+1)]^{N-1-k} \times \\ & \left(F_1(j+1) + \frac{k}{k+1} f_1(j) \right) \Bigg\} \Bigg]. \end{aligned}$$