A New Approach to Image Copy Detection Based on Extended Feature Sets

Jen-Hao Hsiao, Chu-Song Chen, *Member, IEEE*, Lee-Feng Chien, and Ming-Syan Chen, *Fellow, IEEE*

***Abstract*—Conventional image copy detection research concentrates on finding features that are robust enough to resist various kinds of image attacks. However, finding a globally effective feature is difficult and, in many cases, domain dependent. Instead of simply extracting features from copyrighted images directly, we propose a new framework called the *extended feature set* for detecting copies of images. In our approach, virtual prior attacks are applied to copyrighted images to generate novel features, which serve as training data. The copy-detection problem can be solved by learning classifiers from the training data, thus, generated. Our approach can be integrated into existing copy detectors to further improve their performance. Experiment results demonstrate that the proposed approach can substantially enhance the accuracy of copy detection.**

***Index Terms*—Extended feature set (EFS), Gaussian mixture model, image copy detection, ordinal measure, pattern classification, support vector machine.**

I. Introduction

With the growing popularity of the Internet and advances in computer technology, digital images can be distributed ubiquitously. Since it is possible to copy, alter, and distribute digital content easily to a large number of recipients, maintaining intellectual property rights (IPR) in the digital world has become an important issue. Thus, the legal owners of digital content need a way to check whether the images available through a third party originated from their own image collections.

Digital watermarking was the first solution developed to prevent the abuse of digital images. Many digital watermark schemes, such as spectrum watermarks [9], quantization watermarks [8], and blind detection watermarks [39], have been proposed to protect digital images. A digital watermark is basically an identification code that carries information about the copyright owner. It can be invisible and permanently embedded in digital data for copyright protection, ownership verification, and integrity verification. The effectiveness of a watermark-based protection system depends to a large extent on the robustness of the associated digital watermarking method [19], [20]. However, the embedded watermark is not expected to survive under several kinds of attacks. In practice, although many techniques have been proposed, watermark-based frameworks still suffer from robustness problems; consequently, malicious users could remove a watermark via postprocessing.

Recently, the concept of content-based copy detection has been proposed as an alternative means of identifying illegal image copies. The idea is that, instead of hiding additional information in an image to enable copy detection, the image itself can be employed for the same purpose. A content-based copy detection system works as follows: given an image registered by the owner, the system can determine whether near-replicas of the image are available on the Internet or through an unauthorized third party. If it is found that an image is registered (i.e., it belongs to a content owner), but the user does not have the right to use it, the image will be deemed an illegal copy. The suspect image is then sent to the content owner for further identification and a decision about taking legal action against the user.

Content-based copy detection can be used to distinguish illegal copies on its own, or it can complement digital watermark techniques. One way to combine the two methods is to employ a copy detector to find near-replica images initially, and then extract the digital watermarks to confirm ownership.

Some researchers consider that the content-based copy detection is a kind of content-based image retrieval (CBIR) [1], [2], [13], [16], which is widely used to retrieve desired images from a large collection of images. Nevertheless, there is a difference between content-based copy detection and CBIR. An image copy detector searches for near-replicas of an image, whereas CBIR not only retrieves image replicas, but also images that share the same or similar semantics. Copy detection can, thus, be treated as a restricted case of CBIR. However, it is not usually feasible to apply existing CBIR techniques directly to image copy detection, since they may cause a considerable number of false alarms. Therefore, in this paper, we focus on copy detection techniques.

First, we review the literature on copy detection methods. To the best of our knowledge, the work of Chang *et al.* [7] was the first to study this topic. It proposed a near-replica search engine called RIME (Replicated IMage dEtector) for detecting unauthorized copies of images on the Internet. The authors characterized images using wavelets and $C_1C_2C_3$ color space. Subsequently, a clustering technique [6] was developed to improve the efficiency of RIME. Although the method can detect slightly modified images with a high degree of accuracy, it may have difficulty identifying seriously distorted images.

Manuscript received March 30, 2006; revised March 19, 2007. This work was supported in part by NSC96-2422-H-001-001 from the National Science Council, Taiwan. R.O.C. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tamas Sziranyi.

J.-H. Hsiao is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C., and also with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. (e-mail: jenhao@iis.sinica.edu.tw).

C.-S. Chen and L.-F. Chien are with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. (e-mail:song@iis.sinica.edu.tw).

M.-S. Chen is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. (e-mail: mschen@cc.ee.ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.900099

Bhat and Nayar [4] found some defects in correlation-based methods for measuring the distance between images. The typical L_1 and L_2 norms are not robust when an image contains too many outlying pixels. For this reason, they proposed the use of ordinal measures for image matching, whereby an image is partitioned into $m \times n$ equal-size blocks. The average value of each block is then calculated and converted to a rank matrix, which represents the ordinal measures and can be used for similarity measurement. In [17], Kim showed that the use of an ordinal measure of the DCT coefficient is more robust for resisting image modifications and attacks, and employed the MAP (maximum a posteriori) criterion to find an optimal threshold.

In [21] and [37], a dynamic partial function (DPF) was proposed for discovering a better perceptual distance function through mining a large set of visual data. Subsequently, Meng *et al.* [25] developed an enhanced DPF to solve the "one-size-fits-all" problem in DPF. The method uses three schemes to address the image copy detection problem, namely, a thresholding scheme, a sampling scheme, and a weighting scheme. The thresholding scheme allows different numbers of features to be selected in a pairwise manner. Then, the sampling scheme and the weighting scheme are used to substantially improve the detection accuracy. More recently, Qamra and Chang [29] improved the DPF scheme by introducing a locality sensitive hashing (LSH) technique for indexing images.

The use of global image features, as introduced above, may limit the performance of copy detection methods, since only images that are globally similar to the query image will be returned. To resolve this problem, some local region-based algorithms have been proposed. For example, Amsaleg *et al.* [1] and Berrani *et al.* [3] use local descriptors to capture the characteristics of images. They compute many descriptors for each image, where one descriptor corresponds to a region of interest in the image. In [38], Yan *et al.* propose a part-based image copy detector. First, they use a difference of Gaussian (DoG) detector to construct Gaussian pyramids and then search for scale-space extrema (i.e., keypoints) by scanning the image over locations and scales. The keypoints are represented as local descriptors by using PCA-SIFT (principle components analysis on a scale-invariant feature transform). PCA-SIFT extracts a 41×41 pixel patch at the given scale and rotates it to a canonical orientation. The patch is used to generate a compact feature vector by PCA (principle components analysis), which is then employed to construct a distinctive local descriptor for near-duplicate image matching.

Media hashing is another method of content identification and copy detection. The technique differs from conventional cryptographic hashing functions whose outputs change dramatically, even when only one bit of input is changed. However, images with reasonable amounts of distortion are still deemed the same as the original image, so the required hashing function must be less sensitive to image variations. Venkatesan *et al.* [33] proposed an image hash function that converts the traditional hash function to a valid one for copy detection. Their algorithm, which uses randomized signal processing strategies to compress images into random binary strings in a nonreversible way, has been shown to be robust against some image changes. In [24], Mihcak *et al.* used iterative geometric techniques that can tolerate geometric distortion in images. By so doing, their algorithm can withstand slight geometric distortions. More recently, Lu *et al.* [23] proposed a geometry-invariant image hashing scheme that uses mesh-based hash extraction and hash matching for similarity measurement. The proposed method can handle geometric attacks better than conventional media hash techniques. It is worth noting that, although media hashing originated from cryptography, it can still be viewed as a method of discovering a robust feature vector in an image that can tolerate errors caused by attacks.

In this paper, we propose an approach that enhances the detection accuracy, instead of finding attack-invariant image features only in a copyrighted image. The remainder of the paper is organized as follows. In Section II, we present the main concept of our work. In Sections III and IV, we introduce some classifiers incorporated into our framework to solve the copy-detection problem. A comparative evaluation of their respective performances in terms of both detection accuracy and speed is also given. In Section V, a detection cascade is developed to balance the detection accuracy and time, and associated experimental results are detailed to demonstrate its effectiveness. Finally, a discussion and our conclusion are presented in Sections VI and VII, respectively.

II. Main Concept

A. Existing Frameworks

Most copy detection methods emphasize finding good image features for copy detection. We reviewed various kinds of features in Section I. Typically, existing approaches consist of a registration stage and a detection stage. In the registration stage, images are mapped into feature vectors, which are then stored into the database. In the detection stage, the features of a test image are extracted with the same feature-extraction procedures used in the first stage, and then matched with those stored in the registration stage. This is done by computing the distances between the features of the test and the registered images in the feature space, and then applying some thresholds to decide whether the test image is an illegal copy.

In the above scenario, the accuracy of the copy detector depends to a large extent on the robustness of the feature, and on a suitable threshold that can balance false rejection and false acceptance rates. However, although features with possibly high identification power have been introduced, they may not be effective under various kinds of attacks. This reflects a limitation of existing approaches: they lack the ability to exploit useful prior information, such as possible attack models, to boost copy detection performance—even when such information is easy to generate or acquire.

The limitation makes existing approaches vulnerable to malicious attacks. Fig. 1(a) illustrates this phenomenon. In practice, a feature of an attacked image, say $\boldsymbol{A}$, can often be successfully detected, but some others, such as $\boldsymbol{B}$, cannot be detected if they are far away from $\boldsymbol{I}$ in the feature space. However, increasing the acceptance range threshold ε to include the attacked images $\boldsymbol{B}$

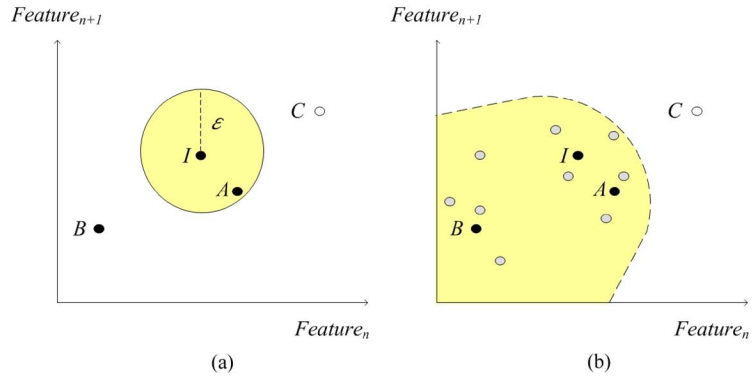

Fig. 1. Let ***I*** denote the feature vector of a copyrighted image, ***A*** and ***B*** be the vectors obtained by applying some attacks to the copyrighted image, and ***C*** be an unrelated feature vector. (a) The radius of the cluster ε denotes the error tolerance for finding copies in the feature space. In practice, an attack on a feature, say ***A***, can often be successfully resisted, but attack on some other feature, such as ***B*** cannot be detected if it is far away from ***I*** in the feature space. However, increasing the threshold of the acceptance range ε to include the attacked image ***B*** could result in a very high false-alarm rate. In this case, ***C*** could be wrongly detected as a copyrighted image. (b) The concept of using EFS to enhance the performance of copy detection, where the gray points are the extended features generated from prior simulated virtual attacks. The boundary between the copyrighted image and unrelated images can be defined more precisely by learning a classifier; thus, the copy detection problem can be solved more effectively.

could result in a high false-alarm rate. In this case, an irrelevant image ***C*** could be falsely detected as a copyrighted image.

In this paper, we propose a new method called the *extended feature set* (EFS) for copy detection. By employing pattern classification techniques, the proposed approach enhances the accuracy of copy-detection, without significantly increasing the detection time.

B. New Scenario

We use an unconventional approach that uses simulated (or virtual) attacks as prior guidance to extract features from a copyrighted image. This information is exploited so that the boundary between a copyrighted image and unrelated images can be defined more accurately. Typical virtual attacks considered in our approach include signal-processing attacks, geometric attacks, and image-compression attacks. By applying the attacks to a copyrighted image, a set of novel images can be generated. Both the copyrighted and the novel images are processed by extracting their features, referred to as the *original* and *extended features*, respectively. Fig. 1(b) shows the concept of extended features in a 2-D space. By using the extended features, we expect to solve the copy-detection problem more effectively.

To train a classifier for copy detection, not only positive examples (which are mainly extended features), but also negative examples are used in our approach. The latter are easy to acquire or generate; for example, they can be collected from the Internet. Note that a registered image can also serve as a negative example of another registered image. Compared to approaches that identify copies based only on the distance between the input image and the original copyrighted image in the feature space, the proposed approach is more effective for copy detection. Our experiment results demonstrate that it generally outperforms conventional methods when the same feature space is employed.

C. Feature Space Employed

To realize the proposed framework, we have to select a feature space onto which an image can be projected. Without loss of generality, we choose the DCT ordinal feature proposed in [17] to build the feature space. In the past, the ordinal measure [4] was widely adopted for applications in image/video retrieval [36] and copy detection [17], [18]. The DCT ordinal feature is particularly suitable for efficient image copy detection over the Internet, since it can be applied to compressed image formats (such as JPEG). However, a limitation of the features generated by the ordinal measure is that they are not robust against geometric attacks. We will show that our approach can effectively boost the performance of the DCT ordinal feature by integrating it into the EFS framework.

Note that, although we use the ordinal feature in this work, our framework is still effective when other features are chosen. Since our approach "increases" the number of features of a copyrighted image by considering modified copies of it generated by simulated attacks, one may wonder whether it is similar to extracting a large set of features (by applying various feature extractors) directly from the copyrighted image. Basically, the two approaches are different because the latter tries to select a set of features that can adequately represent the original (copyrighted) image by using multiple features. However, how to find such a set of features that are invariant to various image attacks remains a key question. To this end, our approach uses different versions of an image generated from prior simulated attacks to learn features with the necessary invariance to cope with image manipulations.

Another distinction is that the number of extended features in our approach can be grown almost infinitely by applying the *priori* simulated attacks. It is easy to find thousands, or even tens of thousands, of extended features for copy-detection by using our framework. In contrast, it is hard to create the same number of useful image features by simply applying various feature extraction methods to a single (i.e., original) image.

In addition, our framework can be used when a set of features is computed for an image. In this case, we can still use different versions of an image (synthesized by mounting attacks on the original image) to generate "even more" features by applying the same feature extraction methods to each version; thus, it

should be possible to learn a better copy detector. Note that the framework proposed in this paper is independent of the selection of feature types. It boosts copy-detection performance by using virtual prior attacks. The use of ordinal features is just one of the techniques used to implement our framework.

D. Learning Methods Employed and Overview of Our Approach

We now present an overview of our approach. To learn the copy detectors, we use three pattern classification methods: the multivariate Gaussian, the Gaussian mixture model (GMM), and the support vector machine (SVM). The first method simply models positive or negative data with a Gaussian distribution. The second models the data by a mixture of multiple Gaussian distributions. The third maps the data into a high-dimensional space (in our framework, a Gaussian kernel function is used) and finds the maximum-margin separation hyperplane in that space. The trained classifiers are then applied to test images not seen in the training stage to evaluate the copy detection performance.

We care about both the accuracy and speed of the copy detector trained. Since it is difficult to model the exact distribution of the training data, we simply try some popular classification methods, namely, SVM and GMM, which have proven effective for solving many pattern classification problems. As will be seen in Section IV-C, both methods improve the accuracy of the original ordinal-measure approach by substantially increasing both the precision rate and the recall rate. However, their classification speeds are slow because multiple evaluations of Gaussian densities are needed, which reduces the detection efficiency. Hence, we employ a two-stage approach that forms a decision cascade. The first stage employs a classifier that is not as effective as the one used in the second stage, but it is faster for classification; and the second stage employs SVM or GMM, which perform better in terms of classification accuracy, but they are slower. The first-stage classifier is constructed by the multivariate Gaussian approach, which simply assumes that both the positive and negative training data can fit a Gaussian distribution. Although this assumption is probably not valid in practice, it can be employed to train a classifier that is computationally efficient.

In the following, we introduce each approach and present the results of employing them for training image copy detectors in our framework.

III. Using the Multivariate Gaussian Classifier in the EFS Framework

We consider the construction of an efficient classifier for image copy detection based on the multivariate Gaussian classifier of two classes in this section. We assume that both the positive and negative examples correspond to Gaussian distributions in which the probability density function is

$$Prob(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-u)^t \Sigma^{-1}(x-u)\right] \quad (1)$$

where d is the feature space dimension, x is a d-dimensional vector in the feature space, and u and Σ denote the $d\times 1$ mean vector and $d \times d$ covariance matrix, respectively.

Let the positive training sample set be $D_P = \{x_0, x_1, \ldots, x_{n-1}\}$, where x_0 is the feature vector of the original copyrighted image, and $x_1, \ldots, x_{n-1}$ are the extended features generated by applying virtual prior attacks to the image. Let $\theta_P = (u_P, \Sigma_P)$ be the unknown mean vector and covariance matrix. As suggested in [11], we estimate the maximum-likelihood (ML) estimation of $\widehat{\theta p}$ by

$$\hat{u}_P = \frac{1}{n}\sum_{i=0}^{n-1} x_i \quad (2)$$

and

$$\hat{\Sigma}_P = \frac{1}{n}\sum_{i=0}^{n-1} (x_i - \hat{u})(x_i - \hat{u})^t. \quad (3)$$

The mean vector and covariance $\mathrm{matrix} = (u_N, \Sigma_N)$ for the negative training set D_N can be estimated in a similar manner.

We assume that the covariance matrix is diagonal (i.e., the elements of a feature vector are statistically independent) for fast computation. For a nondiagonal covariance matrix, the time complexity for density evaluation in (1) is $O(n^2)$, but it is only $O(n)$ when the matrix is diagonal. Note that evaluating the Euclidean distance in the feature space also requires $O(n)$ computations. Thus, the resulting classifier has the same time complexity as that of the original ordinal-measure approach, which finds a copy by calculating the Euclidean distance between two vectors in the feature space. Hence, compared to the original approach we can train an effective copy detector at the expense of only a small increase in detection time.

To determine whether an input image is a copy, the following likelihood ratio L is used:

$$L = \frac{\mathrm{Prob}(z; \hat{\theta}_P)}{\mathrm{Prob}(z; \hat{\theta}_N)} \quad (4)$$

where z is the feature of the input image. The determination is then made by a threshold η. When $L > \eta$, the input is deemed a copy, otherwise it is not.

The decision threshold η can be determined in various ways, such as by the Neyman-Pearson criterion [31]. Instead of simply using a particular threshold for testing, we show the overall performance under various thresholds (cf. Section III-B) to demonstrate the effectiveness of the proposed approach.

In the following, we present the experiment results of the EFS framework when multivariate Gaussian classifiers are used. We first describe the common setup of our experiments in Section III-A, and then report the experiment results in Section III-B. Finally, in Section III-C, we propose a more efficient detector and discuss its performance.

A. Experimental Setup

As mentioned in Section II, we use Kim's approach [17] as the baseline for comparison. In this approach, an input image is divided into 8×8 equal-sized subimages. Only AC coefficients

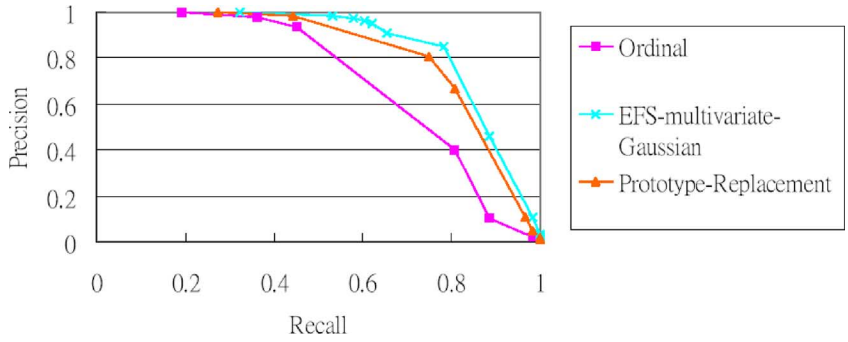

Fig. 2. PR curve for the DCT ordinal measure, EFS-multivariate-Gaussian, and prototype-replacement approaches. Note that the computation time is the same for the ordinal and prototype-replacement approaches.

of the 8×8 DCT coefficients are used to form an ordinal measure. We, thus, generate a 63-dimensional image feature vector.

To build the experimental data set, we collected images from the Corel image CDs and the Internet to form a base set of original images. One hundred images were then randomly selected from the base set to form a copyrighted image set, and the remaining images in the base set served as an irrelevant image set. The latter was further divided into an irrelevant training image set and an irrelevant test image set, which contained 20 000 and 15 000 images, respectively. Note that the sets of training and test images did not overlap.

In the learning stage, each copyrighted image was manipulated with seven different kinds of prior simulated image attacks, namely, noise, blurring, sharpening, cropping, JPEG compression, rescaling, and rotation, which generated 1,679 modified images of the original copyrighted image. The modified images were used to produce extended features (i.e., positive examples), and the irrelevant training images were used as negative examples.[1] Both types of images (i.e., positive and negative examples) were then used to train a classifier based on the multivariate Gaussian approach introduced earlier.

B. Experimental Results for the Multivariate Gaussian Classifier

In the testing stage, we used the one hundred copyrighted images as queries to determine how many modified versions of them could be detected successfully. A standard benchmark, StirMark 4.0 [27], [28], was used to generate novel test data from the one hundred copyrighted images. The image replicas were randomly generated by using the following 13 image attacks from those listed by StirMark 4.0: Convolution filtering (including blurring and sharpening); Cropping (into $20\% \sim 95\%$ sizes); JPEG (with the quality factor ranging from 95% to 5%); Noise adding; Scaling (ranging from 30% to 300%); Rotation (from 1–359 degrees); Median filtering; Affine transformation; Self-similarity (changing color space); Removing lines (frequency from 10 to 100); PSNR (all pixels have had the same values added, ranging from 10 to 100); Rotation+ReScaling (rotation degree from $-30°$ to $30°$); and Rotation+Cropping (rotation degree from $-30°$ to $30°$).

[1]For instance, for the simulated *priori* cropping attacks, we cropped rectangular regions arbitrarily from a copyrighted image. Each region was then divided into 8×8 equal-sized blocks to extract the copyrighted image's ordinal feature.

Note that while some types of attacks (e.g., JPEG, noise adding, scaling, rotation, cropping, and convolution filtering) were trained in advance, some other types (e.g., affine transformation, self-similarities, removing lines, PSNR, median filtering, Rotation+ReScaling, and Rotation+Cropping) were not used in training. Hence, in our experiments, the latter attack types that are unseen in the training stage were included in the testing phase to evaluate the performance of our approach. In addition, for the pretrained types of attacks, different parameters were used to randomly produce new test images, so the training and testing images were totally different.

We generated a total of 124 near-replicas for each copyrighted image: 69 from the pretrained attacks and 55 from unseen attacks; hence, there were 12 400 positive-class images in the test. In addition to the image replicas that served as positive test data, the irrelevant test images described in Section III-A were used as negative test data to evaluate the false alarm rate of our approach, resulting in a total of 27 400 test images.

We use the precision and recall rates to evaluate the copy detection performance. Let r_P be the number of relevant copies correctly assigned to the positive class; let f_P be the number of irrelevant images incorrectly assigned to the positive class; and let r_N be the number of relevant copies incorrectly rejected by the positive class. Then, we can express the precision and recall with r_P, f_P, and r_N as follows:

$$\text{recall} = \frac{r_P}{r_P + r_N} \qquad \text{precision} = \frac{r_P}{r_P + f_P}. \tag{5}$$

The results are summarized in the precision-recall (PR) curve shown in Fig. 2. The curve marked "ordinal" was obtained by using Kim's DCT-ordinal-measure algorithm [17] with a threshold that varies in the feature-space distance. The curve marked "EFS-multivariate-Gaussian" was obtained by using the multivariate Gaussian method with a threshold that varies in the likelihood ratio. As can be seen, our EFS-multivariate-Gaussian approach achieves better precision for all recall rates. Thus, the detection accuracy can be improved by simply adopting a classifier constructed by modeling both classes with Gaussian distributions.

Table I shows the values of the precision/recall-breakpoint (BEP) for both methods, where the BEP point is the point where the precision and recall are equal (or very close) in the PR curve. BEP has been widely used as a performance measure in classification problems. Table I shows that our framework achieves a

TABLE I
BEP AND THE AVERAGE DETECTION TIME OF THE ORDINAL MEASURE, EFS-MULTIVARIATE-GAUSSIAN, AND PROTOTYPE REPLACEMENT APPROACHES; DETECTION TIME CONSISTS OF THE FEATURE-EXTRACTION AND CLASSIFICATION TIMES

Algorithm	BEP Precision	BEP Recall	Avg. detection time (ms)
DCT ordinal measures	62.24%	63.11%	1.1
EFS-Multivariate-Gaussian	85.08%	78.22%	1.28
Prototype-Replacement	80.71%	75.64%	1.1

better performance in terms of BEP than the DCT ordinal measure.

The average detection time, shown in Table I, comprises the time required for feature extraction and the time required for classification. Since the covariance matrix is diagonal, the average detection time only increases slightly (from 1.1 to 1.28 ms).

C. Enhancement of Detection Efficiency and Experiment Results

The above experiments show that the detection accuracy can be improved with only a slight degradation in the detection speed. It would be interesting to know whether we can achieve better detection accuracy than the original (ordinal-feature) approach without any increase in detection time. To this end, we reduce the covariance matrix Σ from an arbitrary diagonal to a special diagonal case, $\Sigma = \sigma^2 I$, where I is the $n \times n$ identity matrix and σ is the standard deviation. It is well known that, in this case, the density evaluation can be simplified by computing the Euclidean distance between the test data and the mean vector $\hat{u}$.

Given a test data set z, instead of computing the likelihood ratio by evaluating the densities, we compute the Euclidean distance directly from z to $\hat{u}_P$, the mean vector estimated for the positive class. Then, we apply a threshold T to this distance to detect copies. In other words, if x_0 is the feature vector of the copyrighted image, the ordinal-feature approach detects copies by computing the Euclidean distance between z and x_0; however, we simply replace x_0 with $\hat{u}_P$ and perform the same detection procedure. Hence, our approach takes exactly the same detection time as the original ordinal-feature approach. Since this approach is equivalent to replacing x_0 in the originally referenced approach by $\hat{u}_P$, we call it the *prototype-replacement* approach.

We used the same positive and negative test data as the previous experiment. Fig. 2 shows the PR-curve obtained by the prototype-replacement approach. From this curve, we observe that the detection accuracy can be enhanced, even when we only replace the copyrighted image feature with the one learned from the positive set. The prototype-replacement approach is more efficient in terms of the detection time (1.1 ms), with the detection accuracy being only slightly lower than that of the EFS-multivariate-Gaussian approach.

In summary, we have shown that, even when simple distribution models and classification rules are used, the proposed EFS framework can improve copy detection accuracy. Furthermore, the increase in computational overhead is limited or even nonexistent.

IV. USING MULTIMODEL CLASSIFIERS IN THE EFS FRAMEWORK

In the previous section, we fitted a Gaussian distribution to both the positive and negative training data. Although the classification performance (in terms of precision and recall) has already been improved, it may not achieve the best performance. In this section, we focus on improving the performance of copy detection when the detection speed is not a major issue. We examine two multimodel approaches for classifier-based image copy detection: the Gaussian mixture model and the support vector machine (SVM). We discuss the models in the following subsections.

A. Mixture of Gaussians

A Gaussian mixture model is defined as

$$f_k(x|\theta) = \sum_{j=1}^{k} w_j g(x|\lambda_j) \tag{6}$$

where $g(x|\lambda_j)$ is a multivariate Gaussian distribution, as defined in (2); $\lambda_j = (u, \Sigma)$ denotes the mean and the covariance matrix of the jth Gaussian component; w_j denotes the weight of the jth component; k denotes the number of components; and $\theta = \{w_j, \lambda_j | j = 1, 2, \ldots, k\}$ is the model parameter set.

To train the GMM model for each class, we apply the expectation–maximization (EM) algorithm, which can converge to a maximum likelihood estimate of θ. In this paper, we used the library in [34] for GMM training.

B. Support Vector Machine (SVM)

In our two-class copy detection problem, SVM [32] can help find the decision boundary with the maximum margin. Given a set of labeled training samples, $S = \{(x_1, y_1), (x_2, y_2), \ldots (x_l, y_l)\}$, $y \in \{1, -1\}$, SVM finds the maximum-margin hyperplane classifier with the boundary $w^T x + b = 0 (w \in R^n, b \in R)$ by solving the following equation:

$$\min_{w,b} \frac{1}{2} w^T w \quad \text{subject to} : y_i((w^T x_i) + b) \geq 1, i = 1, \ldots, l \tag{7}$$

where (w, b) specifies a hyper-plane in the n-dimensional space. To deal with problems that are not linearly separable, SVM maps the input data into a high-dimensional feature space by a transformation Φ, and employs an equivalent kernel operation, $K(x_i, x_j) = \Phi^T(x_i)\Phi(x_j)$, to compute the inner product in the high-dimensional space efficiently. The radial basis function kernel and the polynomial kernel are two of the most commonly used kernels. In our work, we employ the radial basis function kernel

$$K(x_i, x_j) = e^{-\gamma ||x_i - x_j||^2} \tag{8}$$

where γ is a parameter to be adjusted. The decision function is defined as

$$f(x) = \text{Sgn}(w^T \Phi(x) + b) = \text{Sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \tag{9}$$

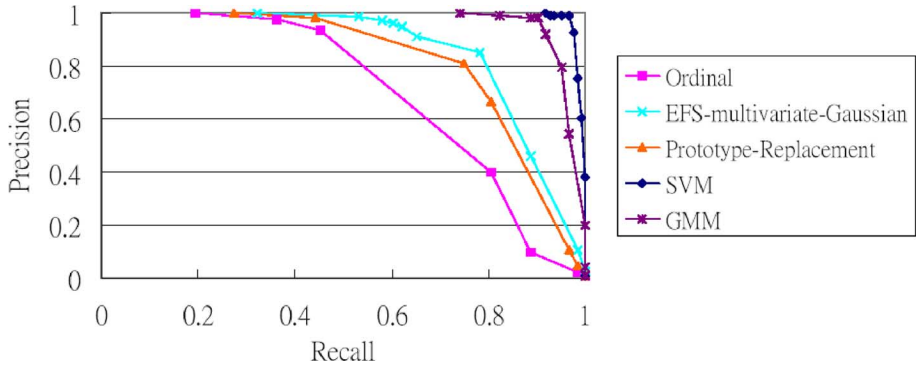

Fig. 3. PR curves for GMM and SVM.

TABLE II
PRECISION AND RECALL RATES OBTAINED USING GMM AND SVM IN THE EFS FRAMEWORK. THE THRESHOLDS OF BOTH METHODS WERE SET ACCORDING TO THE BEP. THE DETECTION TIME IS COMPRISED OF THE FEATURE EXTRACTION AND CLASSIFICATION TIMES

Algorithm	BEP Precision	BEP Recall	Avg. detection time (ms)
DCT ordinal measures	62.24%	63.11%	1.1
GMM	96.56%	93.54%	2.5
SVM	99.27%	96.77%	2.2
2-stage detection cascade	91.20%	91.93%	1.35

Fig. 4. Five color images ($512 * 512$ pixels) used in the smaller-scale experiment: an airplane, a baboon, a selection of fruit, Lena, and peppers.

where the factors α_i are non-negative Lagrange multipliers. To find the maximum margin decision boundary, the following quadratic programming problem must be solved:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \quad y_i(w \bullet \Phi(x_i) + b) \geq 1 - \xi_i \qquad (10)$$

where $\Sigma \xi_i$ is an upper bound of the empirical risk (with $\xi_i \geq 0$ for all $i = 1 \ldots l$) and C is a penalty parameter. In this paper, we use LIBSVM [5] for SVM training and model-parameter selection.

C. Experimental Results

The experimental setup described in Section III-B is also used for the performance evaluation. The selection of the cluster number k is critical in training a GMM [12]. Since we have prior categorical knowledge about the training data, the number of clusters can be set initially as the number of image attack types, which we want to model in advance. To improve the accuracy, the cluster number k can also be assigned automatically by maximizing the logarithm of the likelihood of the Gaussian mixture model on the training samples, and estimated via cross-validation [35]. In our approach, we initially set k as the number of attack types, and continue adding clusters until the log-likelihood either starts to decrease, or keeps on increasing in increments lower than a specific threshold.

To ensure that SVM works well in practice, it is important to select appropriate models in its learning phase. In our work, we use cross validation to select the best model parameters C and γ for classification, as suggested in LIBSVM [5]. The output of SVM, $f(\cdot)$, in (9) is a function with a value of 1 or -1. It can be converted into a probabilistic output. LIBSVM uses the method introduced in [22] to estimate the probability

$$P(y = 1 | f(x)) = \frac{1}{1 + e^{Af(x)+B}}. \qquad (11)$$

where A and B are estimated by minimizing the negative log-likelihood function. Further details can be found in [22].

The results of the above method are summarized by the PR curve shown in Fig. 3. The "GMM" curve is drawn with a threshold that varies in the likelihood ratio, and the "SVM" curve is drawn with a threshold that varies with the probability in (11). Clearly, the multimodel classifiers improve the detection performance substantially. Both the SVM and GMM methods achieve far better precision-recall performances than the original ordinal-feature approach.

Table II shows the BEP values for both methods. SVM performs slightly better than GMM in both detection accuracy and speed. Both SVM and GMM achieve very high recall rates of 96.77% and 93.54%, respectively, while the precision rate is 99.27% for SVM and 96.56% for GMM. The high accuracy rates show that our classification-based EFS framework is very effective for image copy detection.

The above experiment shows the overall performance of our methods. To observe the robustness of our EFS-based approaches against different attacks in more detail, we performed another smaller experiment based only on the five images shown in Fig. 4. The results are summarized in Table III. The thresholds used in the DCT ordinal measures and our EFS methods (SVM and GMM) were set according to the BEP obtained in the above experiment, and the recognition rate (the rate of correctly-classified test data) is shown in Table III. From Table III, we observe that the robustness against geometric attacks (e.g., cropping, rotation, and scaling) is significantly enhanced by our approaches.

In terms of efficiency, the detection times of the SVM and GMM classifiers (2.2 and 2.5 ms, respectively) are approximately twice that of the DCT ordinal measure (1.1 ms). This means that the response time of our multimodel classifiers is roughly double that of the ordinal approach, but the detection

TABLE III
DETAILED RECOGNITION RESULTS OF SVM AND GMM FOR THE FIVE SELECTED IMAGES. THE FORM "ATTACK TYPE $*k$" INDICATES THAT THIS ATTACK TYPE WAS APPLIED k TIMES. FOR EXAMPLE, THE ROTATION ATTACK WAS APPLIED 18 TIMES. THE FORM "m-n-o" INDICATES THAT THE NUMBER OF IMAGE REPLICAS SUCCESSFULLY DETECTED BY SVM, GMM, AND THE ORDINAL APPROACH WERE m, n, AND o, RESPECTIVELY

Pre-learned types of attacks	Testing Item	Airplane	Baboon	Fruits	Lena	Peppers
v	Convolution Filtering (blur or sharp) * 2	2-2-2	2-1-2	2-1-1	2-2-2	2-2-2
v	JPEG * 14	14-14-14	14-14-14	14-14-14	14-14-14	14-14-14
	Median Filtering * 4	4-4-4	4-4-4	4-4-4	4-4-4	4-4-4
v	Noise * 12	12-12-10	12-12-11	11-12-10	12-12-11	12-12-10
	Self-Similarities * 3	3-3-3	3-3-3	3-3-3	3-3-3	3-3-3
	PSNR * 10	10-10-10	10-10-10	10-10-10	10-10-10	10-10-10
v	Scaling * 10	10-10-10	10-10-10	10-10-10	10-10-10	10-10-10
v	Cropping * 13	10-9-2	9-11-2	10-8-2	11-8-3	11-12-2
v	Rotation * 18	17-18-2	18-17-3	17-17-3	16-16-2	18-18-3
	Affine * 8	8-7-6	8-8-7	8-6-6	8-8-7	8-7-6
	Removing Lines * 10	10-10-8	10-10-8	10-10-9	10-10-8	10-10-9
	Rotation+Rescaling * 10	10-8-3	10-10-3	10-9-2	10-10-2	10-10-2
	Rotation+Cropping * 10	10-9-2	10-10-2	10-9-2	10-10-2	10-10-2
Recognition Rate						
Ordinal		61.61%				
SVM		96.94%				
GMM		94.84%				

performance can be improved considerably when SVM or GMM is used. Hence, for situations where such an increase in time is tolerable, our EFS framework can be used to construct classifiers that are highly effective for image copy detection.

V. DETECTION CASCADE

The classification-based approaches presented in Sections III and IV outperform the conventional approach for image copy detection. Specifically, the multivariate Gaussian method is faster with relatively less performance improvement, while SVM and GMM are slower, but the performance improvement is more pronounced. Because they are complementary, we are motivated to develop a detection cascade for image copy detection. The idea is that, the multivariate Gaussian classifier can serve as an initial-level copy detector in the detection cascade to filter out some unrelated images quickly, and the SVM or GMM then acts as a second-level detector to select final candidates.

Fig. 5 shows the block diagram of the proposed detection cascade, which consists of two stages. To construct an efficient detector, we apply the multivariate-Gaussian approach in the first level, and SVM in the second level. We select SVM because its detection speed is faster than that of GMM.

In the first stage, the intermediate threshold of the likelihood ratio in (4) is set as $\eta = 1$. Note that this is just a setting to perform our experiments, and can be modified in practice by considering different precision/recall or false-detection/false-alarm tradeoffs. Then, the input data that lies in a neighborhood of the decision boundary of the first stage is sent to the second stage for additional verification. The images whose likelihood ratios are within the interval $[1-r, 1+r]$, where r is a positive value specifying the ambiguous range, are sent to the second stage for further examination; otherwise, the first-stage decisions serve as the outputs directly.

To evaluate the performance of the 2-stage detection cascade, the experiments described in Section III-B were also used for the detection-cascade approach. The selection of an ambiguous

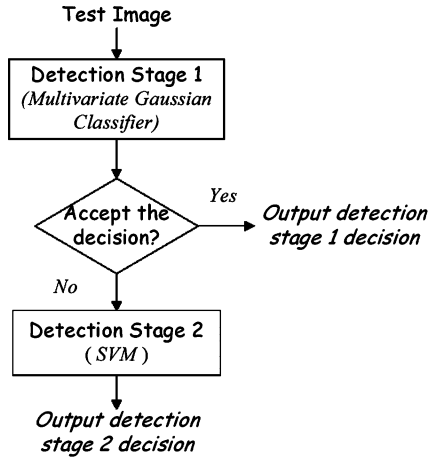

Fig. 5. Block diagram of the 2-stage detection cascade.

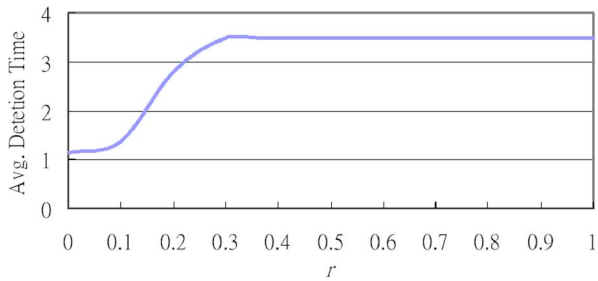

Fig. 6. Average detection time of the 2-stage detection cascade obtained by varying the ambiguous range r.

range r is a tradeoff between the detection accuracy and the speed. A larger r leads to a better recognition rate, but results in a slower detector. Fig. 6 shows the average detection time of the 2-stage detection cascade derived by applying different r values to the training data. We observe that the detection time increases sharply when $r > 0.1$. Hence, we choose $r = 0.1$ to implement the 2-stage detection-cascade approach.

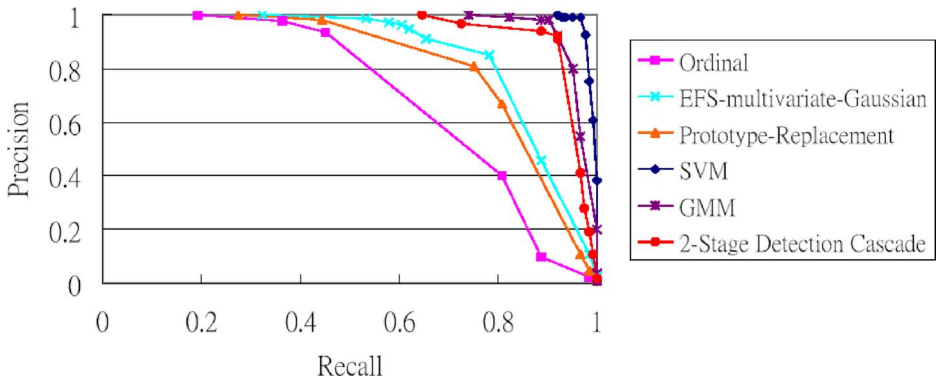

Fig. 7. PR curve for the 2-stage detection cascade/.

The results are summarized in the PR curves shown in Fig. 7. The curve marked "2-Stage Detection Cascade" is obtained by using a threshold that varies in the probability ratio (11) of the second-stage classifier (i.e., SVM). The 2-stage detection cascade achieves better detection accuracy than the multivariate-Gaussian and the prototype-replacement approaches.

Table II shows the BEP performance of the 2-stage detection cascade approach. The detection cascade approach affords a suitable tradeoff between the discriminative ability and computational efficiency. A simple model, such as the multivariate Gaussian classifier is fast (1.28 ms), but it has less discriminative ability (78.22% recall and 85.08% precision at BEP). Multimodel approaches, such as SVM, are highly accurate (96.77% recall and 99.27% precision at BEP), but slower (2.2 ms). However, the detection cascade approach is not only very accurate (91.2% recall and 91.93% precision at BEP); it is also fast (1.35 ms).

In summary, our EFS framework can be used to construct various classifiers that are suitable for different accuracy and/or speed requirements. These approaches can fulfill different application needs and provide generally better detection accuracy than the conventional approach.

VI. Discussion

We now consider the rationale behind EFS. In fact, learning a classifier based on precollected training data has been studied in several general CBIR problems [10], [15]; however, to the best of our knowledge, similar ideas have not been applied to image copy detection previously. In this paper, we have demonstrated that the learning-based approach is suitable for copy detection. We believe the reason is that, unlike the general CBIR, copy detection aims to find near-duplicates, instead of images that have the same semantic meaning.

More specifically, from a user's perspective, an ideal CBIR system should be capable of semantic retrieval, but a copy detector only identifies near duplicates that can tolerate some modification, without needing to include all the images that are similar at the semantic level. Note that, for general CBIR, it is almost impossible to collect all the relevant images for prelearning the general semantic meaning of the content. However, for copy detection, the positive training examples can be obtained in a "generative" manner, and, thus, a relatively more thorough training set can be precollected. This makes the classification-based framework more suitable for solving the image copy detection problem.

VII. Conclusion

In this paper, we have proposed a new and effective scheme that can detect unauthorized copies of images by employing learning techniques. We have presented a general framework that exploits prior information (generated from prior simulated attacks) so that the boundary between the copyrighted and unrelated images can be defined more accurately through classifier learning. Although the image copy detection problem is essentially a pattern classification problem, to the best of our knowledge, no previous works have employed classifier training algorithms to solve it. Experiment results obtained from benchmark attacks confirm the efficacy of the proposed method.

Instead of dealing directly with the feature selection problem, which is hard to solve and domain dependent, the proposed EFS framework addresses the copy detection problem by using prior simulated attacks. This technique enhances the detection accuracy by generating features with the necessary invariance to resist various types of image manipulation. In addition, we construct a 2-stage detection cascade to balance the detection time and classification performance. Although we validate the effectiveness of the proposed approach by using ordinal features and multivariate-Gaussian/SVM/GMM methods, it is independent of the feature types or the classifier learning methods selected. Thus, it can be easily integrated into existing copy detectors to improve their performance.

Extending the proposed approach to a detector that can identify different types of attack is a topic that merits further study. We believe that such an extension could be achieved by replacing the two-class problem with a multiclass one. We will address this issue in our future work.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments which have led to improvements of this paper. They would also like to thank Mr. P. Dunne for his work on polishing the writing.

References

[1] L. Amsaleg and P. Gros, "Content-based retrieval using local descriptors: Problems and issues from a database perspective," *Pattern Anal. Appl.*, vol. 4, pp. 108–124, 2001.

[2] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Jain, and C. Shu, "The virage image search engine: An open framework for image management," in *Proc. SPIE: Storage and Retrieval for Still Image and Video Database IV*, 1996, vol. 2670, pp. 76–87.
[3] S. A. Berrani, L. Amsaleg, and P. Gros, "Robust content-based image searches for copyright protection," in *Proc. ACM Int. Workshop on Multimedia Databases*, 2003, pp. 70–77.
[4] D. N. Bhat and S. K. Nayar, "Ordinal measures for image correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 415–423, Apr. 1998.
[5] C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines 2001 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[6] E. Y. Chang, C. Li, J.-Z. Wang, P. Mork, and G. Wiederhold, "Searching near-replicas of images via clustering," in *Proc. SPIE: Multimedia Storage and Archiving Systems IV*, 1999, vol. 3846, pp. 281–92.
[7] E. Y. Chang, J.-Z. Wang, C. Li, and G. Wiederhold, "RIME: A replicated image detector for the world-wide-web," in *Proc. SPIE: Multimedia Storage and Archiving Systems III*, 1998, vol. 3527, pp. 58–67.
[8] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1423–1443, May 2001.
[9] I. J. Cox, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.
[10] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, Jan. 2000.
[11] R. O. Duda *et al., Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2001, pp. 88–89.
[12] M. A. F. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.
[13] M. Filickner, H. Sawhney, W. Niblack, J. Ashley, W. Huang, B. Dom, M. Gorkani, J. Hafine, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content—The QBIC system," *IEEE Computer*, vol. 28, no. 9, pp. 23–32, Sep. 1995.
[14] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," Tech. Rep. 319 Dept. Statist., Univ. Washington, Seattle, 1998.
[15] D. Hoiem, R. Sukthankar, H. Schneiderman, and L. Huston, "Object-based image retrieval using the statistical structure of images," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 490–497.
[16] M. L. Kherfi, D. Ziou, and A. Bernardi, "Image retrieval from the world wide web: Issues, techniques, and systems," *ACM Comput. Surv.*, vol. 36, pp. 35–67, 2004.
[17] C. Kim, "Content-based image copy detection," *Signal Process.: Image Commun.*, vol. 18, pp. 169–184, 2003.
[18] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 127–132, Jan. 2005.
[19] M. Kutter and F. Petitcolas, "A fair benchmark for image watermarking systems," in *Proc. SPIE: Security and Watermarking of Multimedia Contents*, 1999, vol. 3657, pp. 58–67.
[20] S.-J. Lee and S.-H. Jung, "A survey of watermarking techniques applied to multimedia," in *Proc. IEEE Int. Symp. Industrial Electronics*, 2001, vol. 1, pp. 272–277.
[21] B. Li, E. Y. Chang, and C.-T. Wu, "DPF – A perceptual distance function for image retrieval," in *Proc. Int. Conf. Image Processing*, 2002, vol. 2, pp. 597–600.
[22] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," Tech. Rep. Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, 2003.
[23] C.-S. Lu and C.-Y. Hsu, "Geometric distortion-resilient image hashing scheme and its applications on copy detection and authentication," *ACM Multimedia Syst. J.*, vol. 11, pp. 159–173, 2005.
[24] M. K. Mihcak and R. Venkatesan, "New iterative geometric methods for robust perceptual image hashing," in *Proc. ACM Workshop on Security and Privacy in Digital Rights Management*, 2001, vol. 2320, pp. 13–21.
[25] Y. Meng, E. Y. Chang, and B. Li, "Enhancing DPF for near-replica image recognition," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2003, vol. 2, pp. 416–23.
[26] N. Lachiche and P. A. Flach, "Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves," in *Proc. Int. Conf. Machine Learning*, 2003, pp. 416–423.
[27] F. Petitcolas, "Watermarking schemes evaluation," *IEEE Signal Process. Mag.*, vol. 17, pp. 58–64, 2000.
[28] F. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," in *Proc. Int. Workshop on Information Hiding*, 1998, vol. 1575, LNCS, pp. 219–239.
[29] A. Qamra, Y. Meng, and E. Y. Chang, "Enhanced perceptual distance functions and indexing for image replica recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 379–391, Mar. 2005.
[30] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller, "Constructing boosting algorithms for SVMs: An application to one-class classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1184–1197, Sep. 2002.
[31] C. Scott and R. Nowak, "A Neyman-Pearson approach to statistical learning," *IEEE Trans. Inf. Theory IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 51, no. 1, pp. 3806–3819, Jan. 2005.
[32] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[33] R. Venkatesan, S.-M. Koon, M. H. Jakubowski, and P. Moulin, "Robust image hashing," in *Proc. Inf. Conf. Image Processing*, 2000, vol. 3, pp. 664–666.
[34] *Weka: Data Mining Software in Java* [Online]. Available: http://www.cs.waikato.ac.nz/~ml/weka/index.html
[35] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second ed. San Mateo, CA: Morgan Kaufmann, 2000, pp. 296–297.
[36] H. Wu, H. Lu, and S. Ma, "A practical SVM-based algorithm for ordinal regression in image retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 612–621.
[37] Y.-L. Wu, C.-W. Chang, W.-C. Lai, K.-T. Cheng, and E. Y. Chang, "Multimedia web services for content filtering, searching, and digital rights management," in *Proc. Inf. Conf. Information, Communications and Signal Processing, and 4th Pacific Rim Conf. Multimedia*, 2003, vol. 1, pp. 191–196.
[38] K. Yan, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and subimage retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2004, pp. 869–876.
[39] W. Zeng and B. Liu, "A statistical watermark detection technique without using original images for resolving rightful ownerships of digital images," *IEEE Trans. Image Process. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 11, pp. 1534–1548, Nov. 1999.

Jen-Hao Hsiao received the M.S. degree in computer science from SooChow University, Taipei, Taiwan, R.O.C., in 2002. He is currently pursuing the Ph.D. degree at the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan.

His research interests include content-based image retrieval/detection, machine learning, and digital rights management.

Chu-Song Chen (S'94–M'97) received the B.S. degree in control engineering from National Chiao-Tung University, Hsing-Chu, Taiwan, R.O.C., in 1989, and the M.S. and Ph.D. degrees from the Department of Computer Science and Information Engineering, National Taiwan University (NTU), Taipei, in 1991 and 1996, respectively.

He is now an Associate Research Fellow of the Institute of Information Science, Academia Sinica, Taiwan, and also an Adjunct Associate Professor of the Graduate Institute of Networking and Multimedia, NTU. His research interests include pattern recognition, computer vision, signal/image processing, and multimedia. He has published more than 70 technical papers, and has received the outstanding paper awards of IPPR in 1997, 2001, and 2005.

Dr. Chen serves as the Secretary General of the Image Processing and Pattern Recognition (IPPR) Society, Taiwan, since 2007, which is one of the societies of the International Association of Pattern Recognition (IAPR).

Lee-Feng Chien received the Ph.D. degree in computer science from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1991.

Since 1993, he has been with the Institute of Information Science, Academia Sinica, Taiwan, where he is currently a Research Fellow. His research interests include information retrieval, natural language processing, spoken language processing, and web mining. At present, he is on leave, working with Google as Engineering Director of the Taiwan R&D center.

Ming-Syan Chen (S'88–M'98–SM'93–F'03) received the B.S. degree in electrical engineering from the National Taiwan University (NTU), Taipei, Taiwan, R.O.C., and the M.S. and Ph.D. degrees in computer, information, and control engineering from The University of Michigan, Ann Arbor, in 1985 and 1988, respectively.

He is currently a Distinguished Professor jointly appointed by the Electrical Engineering Department, Computer Science and Information Engineering Department, and also the Graduate Institute of Communication Engineering at NTU. He was a research staff member at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, from 1988 to 1996. He holds, or has applied for, 18 U.S. patents and seven R.O.C. patents in his research areas. His research interests include database systems, data mining, mobile computing systems, and multimedia networking, and he has published more than 240 papers in his research areas.

Dr. Chen is a Fellow of ACM. In addition to serving as program chairs/vice-chairs and keynote/tutorial speakers at many international conferences, he was an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDEGE AND DATA ENGINEERING, as well as the *Journal of Information Science and Engineering*. He is currently on the editorial board of the *Very Large Data Base Journal*, the *Knowledge and Information Systems Journal*, and the *International Journal of Electrical Engineering*. He was a Distinguished Visitor of the IEEE Computer Society for Asia-Pacific from 1998 to 2000 and from 2005 to 2007. He is a recipient of the National Science Council Distinguished Research Award, the Pan Wen Yuan Distinguished Research Award, the Teco Award, an Honorary Medal of Information, and the K.-T. Li Research Breakthrough Award for his research work, as well as the Outstanding Innovation Award from IBM Corporate for his contribution to a major database product. He has also received numerous awards for his research, teaching, inventions, and patent applications.