Recognition of Adult Images, Videos, and Web Page Bags

WEIMING HU, HAIQIANG ZUO, OU WU, and YUNFEI CHEN, Institute of Automation, Chinese Academy of Sciences
ZHONGFEI ZHANG, State University of New York
DAVID SUTER, University of Adelaide

In this article, we develop an integrated adult-content recognition system which can detect adult images, adult videos, and adult Web page bags, where a Web page bag consists of a Web page and a predefined number of Web pages linked to it through hyperlinks. In our adult image-recognition algorithm, we model skin patches rather than skin pixels, resulting in better results than state-of-the-art algorithms which model skin pixels. In our adult video-recognition algorithm, information from the accompanying audio section around an image in an adult video is used to obtain a prior classification of the image. The algorithm achieves a better performance than the ones which use image information alone or audio information alone. The adult Web page bag recognition is carried out using multi-instance learning based on the combination of classifying texts, images and videos in Web pages. Both the speed and the accuracy for recognizing the Web adult content are increased, in contrast to recognizing Web pages one-by-one.

Categories and Subject Descriptors: H.5.1 [Information Interfaces and Presentation]: Multimedia Information System—Evaluation/methodology; H.5.2 [Information Interfaces and Presentation]: User Interface—User-centered design

1. INTRODUCTION

With the explosive growth of the World Wide Web (WWW), there are more and more Web sites providing information often considered offensive and obscene. Thus there is much interest in filtering such information. For example, in 2005, the European Parliament launched a large program called “Safer Use of the Internet”, particularly for young people. In 2009, The Department of Broadband, Communications and the Digital Economy of Australia initialized a Web content filtering system which aims to keep children from adult and other content considered harmful. In this article, we focus on the recognition of Web adult content.

This work was partially supported by the National Science Foundation of China (NSFC) under Grant Nos. 60825204, 60935002. Author's address (corresponding author): W. Hu, Institute of Automation, Chinese Academy of Sciences; email: wmhu@nlpr.ia.ac.cn.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2011 ACM 1551-6857/2011/10-ART28 $10.00
DOI 10.1145/2037676.2037685 http://doi.acm.org/10.1145/2037676.2037685
1.1 Related Work

Text classification has been used to recognize Web adult information. Du et al. [2003] proposed a Web filtering system that uses a text classification algorithm to classify web pages into adult and non-adult pages. Lee et al. [2005] presented a bilingual Web page categorization engine that can determine whether a Web page contains adult content. Ho and Watters [2004] proposed a method of finding differences between adult and non-adult Web pages based on structural and statistical analysis, resulting in an effective Bayesian classification between adult and non-adult texts.

In contrast to adult texts, adult images are considered to be more influential, because image information is much more rapidly perceived and the graphic effects are often more shocking and disturbing to people. There have been many algorithms proposed for recognizing adult images. For example, Forsyth and Fleck [1999] devised a system relating different parts of the human body to their geometric constraints in order to detect naked people. Ioffe and Forsyth [2001] extended Forsyth and Fleck [1999] by using a new sampling method to learn the geometric relations between human body parts. Wang et al. [1998] developed a wavelet analysis-based method to recognize adult images. Jiao et al. [2001] used color coherence vectors and color histograms to identify the images which contain naked people. Liang et al. [2004] extracted multi-features from skin blobs and utilize them as inputs to a multilayer perceptron neural network classifier, which was used to detect adult images. Arentz and Olstad [2004] extracted features from skin regions, based on color, texture, contour, placement, and relative size information for each region in an image and calculated the probability that the image is adult in nature. Zheng et al. [2004] adopted a maximum entropy model for skin detection. A multilayer perception was used to identify adult images. Shih et al. [2007] used the MPEG-7’s color, texture, and shape features to retrieve a number of images most similar to the test image from the image database. If the retrieved images contain more than a predefined number of adult images, the test image is identified as adult. Yang [2006] constructed a real-time image-understanding platform based on a physical linguistic network and cognitive features and applied this platform to detect adult images. Deselaers et al. [2008] used the bag-of-visual-words model to classify images into different categories of adult content. Lienhart and Hauke [2009] evaluated the use of probabilistic latent semantic analysis for detecting adult images. Rowley et al. [2006] presented a method for detecting adult images from a large fraction of Web images. For each image, a number of summary features were computed and a support vector machine was used for classification.

Adult videos are often considered to be more influential than adult images, due to their increased realism and ability to portray actions in detail. There have been several attempts to build adult video detection systems. Han et al. [2005] classified key frames in a video using MPEG-7 visual descriptors. The results of the key frame classification were used to determine whether the video is adult. Lee et al. [2009] statistically estimated the degree to which the video could adult by using single frame-based features and group-of-frame features. Lopes et al. [2009] used bag-of-visual features, which are obtained by adding color information to the original scale invariant feature transform (SIFT), to classify adult videos. Endeshaw et al. [2008] proposed a fast method for detecting adult video content using repetitive motion analysis rather than skin detection. Jansohn et al. [2009] combined conventional key frame-based methods with a statistical analysis of MPEG-4 motion vectors to detect adult videos.

Some researchers combine text information and image information, or combine information from multi-images to recognize Web adult content. Wang et al. [1998] classified a Web site as adult based on the contents of the images. If more than a certain number of images sampled from the Web site are found to be adult, then the site is considered to be adult. Arentz and Olstad [2004] ran an image-based classifier on the whole collection of the images in a Web site and then set up a histogram of images with in order regards to the “adult log-likelihood” for each image to recognize adult Web sites.
Hammami and Chahir [2006] combined textual and structural content-based analysis with skin color-related visual content-based analysis in a series to recognize Web adult information. Jones and Rehg [2002] combined the adult image detector with a text-based detector which deals with the text around an image on a page to determine whether the image is adult. In our previous work [Hu et al. 2007], we combined the results of classifying the images and the texts in a Web page to determine whether the Web page is adult.

1.2 Our Work

This article describes an integrated adult-content recognition system which can handle adult images, adult videos, and adult Web pages. Figure 1 gives an overview of the system. A predefined number of Web pages are selected from a Web page set that is composed of a Web page and the pages linked to it to form a Web page bag. We treat a Web page bag as a bag in multi-instance learning (MIL), and its Web pages are treated as instances in the bag. For each Web page in a bag, the different analysis strategies should be designed specifically to deal, respectively, with texts, images, and videos. We use text processing to extract text features of a Web page, we use the results of classifying images in the Web page to extract its images features, and we use the results of classifying its videos to extract its video features. Next, the text features, the image features, and the video features are concatenated to form the feature vector of the Web page. The feature vectors of the Web pages in the sample bags are used to construct the MIL-based classifier, which is used to classify test Web page bags.

Our system is original in the following ways.

—Skin detection based on skin patch modeling. Most existing skin detection algorithms are pixel-based, classifying each pixel as skin or non-skin individually. Current patch-based skin detection algorithms model skin pixels and detect skin pixels considering patch information, such as the relation between pixel values in a patch. In this article, we directly model skin patches by representing relations between values of pixels in a patch using a covariance descriptor and directly classify test patches as skin or non-skin. Experimental results show that our algorithm outperforms many state-of-the-art skin detection algorithms.

—Global and local combination-based adult image recognition. We extract features from adult images using the idea of going from “global” to “local”. Several high-level features, such as roundness and irregularity of the largest skin region and the results of rapid detection of faces, breasts, and vulvas are introduced into the recognition of adult images.
Recognition of Web adult videos by fusing image information and audio information. We propose an algorithm for recognizing adult videos based on the prior assumption that image information and the audio information in an adult video express the adult theme. In the algorithm, local information from the audio section associated with an image in a video is used to obtain a prior classification of the image. Then, the results of classifying the images and the audio sections in the video are combined to classify the video as a whole.

Recognition of adult Web page bags using MIL. Current research on the recognition of Web adult information treats individual adult Web pages as objects to be handled. In this article, we recognize adult Web page bags using MIL. Bayesian-kNN (nearest neighbor), citation-kNN, mi-SVM, and MISVM are, respectively, applied to carry out this MIL task using the results of classifying texts, images, and videos in Web pages. A comparison between the results of these four MIL algorithms is made. In contrast to classifying Web pages one-by-one, classifying Web page bags not only reduces the runtime, but also increases the recognition accuracy.

The remainder of this article is organized as follows. Section 2 describes our algorithm for recognizing Web adult images. Section 3 presents our algorithm for recognizing adult videos. Section 4 covers our algorithm for recognizing adult Web pages. Section 5 demonstrates experimental results. Section 6 summarizes the article.

2. WEB ADULT IMAGE RECOGNITION

The extraction of image features for a Web page depends on recognition of adult images in the Web page. Our Web adult image recognition algorithm consists of patch modeling-based skin detection [Zuo et al. 2010], global and local feature-based adult image recognition, and hierarchical handling of web adult images.

2.1 Patch Modeling-Based Skin Detection

The orthogonal YCrCb color space is used to detect skin colors, where Y is the luminance (or intensity) component and Cr and Cb are the red-difference and blue-difference chrominance components. An image is split into patches. For each patch, a feature vector is extracted. A random forest-based classifier is trained using skin patch samples, and the trained classifier is used to detect skin patches.

2.1.1 Feature extraction. Given a skin pixel patch, we extract the color distribution features and the features that describe relations between values of pixels in the patch.

Relations between values of pixels in a patch are modeled using the covariance descriptor [Tuzel et al. 2008]. For each pixel at \((x, y)\), we extract the intensity \(I(x, y)\) and the norms of the first- and second-order derivatives \(\partial I(x, y)/\partial x, \partial I(x, y)/\partial y, \partial^2 I(x, y)/\partial x^2, \partial^2 I(x, y)/\partial y^2\) of the intensity in both the \(x\) and \(y\) directions to map the pixel to a 5-dimensional vector \(\mathbf{z}(x, y)\). Then, a patch is represented by a \(5 \times 5\) symmetric covariance matrix \(C\):

\[
C = \frac{1}{S-1} \sum_{i=1}^{S} (z_i - \mu)(z_i - \mu)^T,
\]

where \(S\) is the number of pixels in the patch and \(\mu\) is the mean of the vectors of the pixels in the patch. The covariance matrix has scale and rotation invariance. When the patch is rectangular, integral images are employed to accelerate the computation for obtaining the covariance matrix [Tuzel et al. 2008]. Nonsingular covariance matrices can be formulated as a connected Riemannian manifold. Let \(I_5\) be a \(5 \times 5\) identity matrix. Singular value decomposition (SVD) for \(C\) i.e. \(C = U \Sigma U^T\) produces the orthonormal matrix \(U\) and the diagonal matrix \(\Sigma = \text{Diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)\), where \(\{\lambda_i\}_{i=1}^{5}\) are the
singular values of \( C \). The matrix logarithm of \( C \) is defined as:

\[
\log(C) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (C - I_5)^k
= U \cdot \text{diag}(\log(\lambda_1), \log(\lambda_2), \ldots, \log(\lambda_5)) \cdot U^T
\]  

(2)

The matrix logarithm operation achieves the tangent space projection, which maps covariance matrices into the tangent space at the identity element of the connected Riemannian manifold. This tangent space is a vector space in which regular machine learning techniques can be used. Let \( X \) be the matrix obtained by projecting \( C \) into the tangent space. We unwrap \( X \) into a vector \( x \). As \( X \) is a symmetric matrix, only the elements of its upper triangular matrix are utilized in \( x \) for saving the space and the computational time. Then, the unwrapped vector \( x \) is represented as \( x = (X_{1,1}, \sqrt{2}X_{1,2}, \ldots, \sqrt{2}X_{1,5}, X_{2,2}, \sqrt{2}X_{2,3}, \ldots, \sqrt{2}X_{2,5}, \ldots, X_{5,5}) \), where the coefficient for nondiagonal elements is \( \sqrt{2} \) (to ensure that the distance between any two symmetric matrices is equal to the distance between the corresponding unwrapped vectors). The 15 features in \( x \) describe the relations between values of pixels in a patch.

To represent the color distribution of a patch, we compute the means and standard deviations of the \( C_r \) and \( C_b \) components of the pixels in the patch to form 4 features. Then, we concatenate these 4 features and the 15 features in \( x \) into a 19-dimensional feature vector to represent the patch.

2.1.2 Classifier. The random forest-based classifier [Breiman, 2001] is selected to model skin patches since it is accurate for classification and the model it uses is easily understood. The random forest is composed of a number of decision trees. To train a decision tree, a sample subset is randomly selected from the training samples using the putting-back strategy based on the bootstrap process where the size of the subset is kept the same for all the decision trees. When each node in a decision tree is split, the optimal feature is selected from a subset of features, which are randomly selected from the whole feature set. The output of the random forest-based classifier is the class label that receives the majority votes of all the decision trees in the forest.

2.1.3 Patch-based skin detection. We generate patches from an image and extract their features to determine whether they are skin or non-skin ones. We develop two types of patch generation methods, regular patch-based and irregular patch-based.

(1) Regular patches. These nonoverlapping patches are generated by sliding a rectangular window from left-to-right and from top-to-bottom. Initially, the patch size is set to 32 \( \times \) 32 pixels. If the sum of the standard deviations of the \( C_r \) and \( C_b \) components in a patch exceeds a predefined threshold, the patch is split into 4 patches with 16 \( \times \) 16 pixels. This process is repeated, if necessary, until the patch size is equal to 4 \( \times \) 4 pixels. The merit of regular patches is that integral images can be used to speed up the process of feature extraction. The limitation is that the contours of the detected skin regions can be zigzagged.

(2) Irregular patches. These patches are generated using a standard image segmentation algorithm, such as the algorithm in Felzenszwalb and Huttenlocher [2004]. The merit of irregular patches is that “natural” skin regions can be obtained. The limitation is that the integral images cannot be used to compute covariance matrices for nonrectangular patches, leading to a comparatively high computational complexity.

2.2 Adult Image Recognition

On the basis of skin detection results, we extract recognition features from images using the idea of going from “global” to “local”: The extracted features are composed of global image features and global and local human body features.
Global image features. We use the following 14 features to represent the global property of an image.

—Image aspect ratio.
—Image entropy, which used to distinguish between natural images and artificial images.
—Proportion of skin regions in the whole image.
—Number of skin regions.
—0-3-order spatial moments of skin regions (These 10 features have the property of invariance to rotation, translation, and scale)

Global human body features. We extract the following 13 features from the largest skin region and its fit ellipse to represent the global human body property.

—Ratio of the area of the largest skin region to the skin area in the whole image.
—Distance from the center of the fit ellipse to the center of the image.
—Length ratio of the two axes of the fit ellipse.
—Tilt angle of the fit ellipse.
—Hu's invariant moments of the largest skin region: these 7 features have the property of invariance to shapes.
—Roundness and irregularity [Aragon et al. 2007] of the largest skin region (These two features are obtained by improving the Fourier descriptor to implement a rapid contour shape representation; and we introduce these two new features to describe the shape of the contour of the largest skin region.)

Local human body features. We apply a rapid object detection technique to extract local human body features. In particular, we employ the boosted cascade of Haar-like features (including edge, line, center-surround, and tilt features) in Lienhart and Maydt [2002] and Viola and Jones [2001] to rapidly detect the local parts: faces, breasts, and vulvas. For each object, one classifier is trained using the positive samples with the detected object and the negative samples without the detected object. The detected results are refined by filtering out the detected objects with regions of only a few skin pixels. The following 4 features are used to represent human body's local property.

—Number of detected faces
—Number of detected breasts
—Number of detected vulvas
—Area ratio of skin pixels in the detected faces to skin pixels in the whole image.

After feature vectors are extracted from adult sample images and non-adult sample images, a random forest-based classifier is trained for classifying test images.

2.3 Hierarchical Handling of Web Adult Images

As Web images are mainly non-adult, we design a fast classifier to quickly filter out non-adult images. The remaining images are classified using the adult-image recognition algorithm just described. This hierarchical handling of web adult images greatly reduces the time cost for the Web adult content filtering. The fast classifier is based on the determination of human existence in images.

Current investigations [Ioffe and Forsyth 1999; Dalal and Triggs 2005] on “human mining” in images mainly focused on face detection and human body detection, and they are inapplicable to Web adult images in that Web adult images have great variances in scenes and in human gestures. In this article,
we judge human existence in an image by the clues for existence of human skin. If there is skin, it is possible that there is human existence in the image. The larger the proportion of the skin area, the more probable it is that there is human existence in the image. We note that there are many images that contain people but contain very small skin areas. Such images are not relevant to adult image recognition. It is not sufficient to determine human existence in an image by only using skin information. For a Web image, three types of supporting evidences of human existence can be calculated in terms of probabilities: the image, its surrounding text, and the title of the Web page. The final probability of human existence is obtained by the fusion of these three probabilities (the algorithm is described in the appendix). If the final probability of human existence is larger than a threshold, then the adult-image recognition algorithm is used to further classify this image.

3. RECOGNITION OF ADULT VIDEOS

Extraction of video features for a Web page depends on recognition of the adult videos in the Web page. In adult videos, the image frames and the audio waveform are closely related and provide the consistent content. Fusion of the available information in the audio tracks and in the image frames is a reliable way to recognize adult videos since such fusion eliminates ambiguities possibly existing in individual modalities.

In this article, we propose an adult-video recognition algorithm based on the fact that the images and the accompanying audio in an adult video express the same theme, pornography. In an adult video there must be at least one clip in which almost all the frames contain adult content. This clip is called an adult clip. A video is divided into a number of clips with small sizes (a few dozen frames), and the proposed algorithm is applied to recognize the clips one-by-one. Once one of the clips is classified as adult, the whole video is classified as adult. In order to increase the running speed of the recognition process, only frames at predefined intervals in a clip are selected for the recognition process.

There is a priori knowledge in adult clips: The information in the audio accompanying an adult video provides a prior classification of the video or prior classifications of the images in the video. For example, the local information in the audio section around an image in an adult clip provides a prior classification of that image. Then, the task of adult clip recognition is to determine whether a given clip is adult or non-adult according to the results of classifying the images and the audio sections around these images in the clip. Clips are classified into adult category \( c_1 \) and non-adult category \( c_2 \). Correspondingly, adult images belong to category \( c_1 \), and other images belong to category \( c_2 \). As the output of the adult-image recognition algorithm is a decision about whether an image belongs to \( c_1 \) or \( c_2 \), the result of classifying image \( i \) can be represented as \( \mu_i \), where \( \mu_i \in \{c_1, c_2\} \). Then, the results that all the images in a clip are classified using the adult-image recognition algorithm are represented as the vector \( R = (\mu_1, \mu_2, \ldots, \mu_i, \ldots, \mu_N) \), where \( N \) is the number of the images in the clip. Our task is to decide whether a given clip belongs to \( c_1 \) or \( c_2 \) using \( R \).

Images in a clip are classified independently. Thus, the classification results of the images in a clip are independent of each other. Then, the following equations are obtained:

\[
p(R|c_1) = p(\mu_1, \mu_2, \ldots, \mu_N|c_1) = \prod_{i=1}^{N} p(\mu_i|c_1),
\]

\[
p(R|c_2) = p(\mu_1, \mu_2, \ldots, \mu_N|c_2) = \prod_{i=1}^{N} p(\mu_i|c_2).\]

The probability \( p(\mu_i|c_1) \) or \( p(\mu_i|c_2) \) represents the probability that image \( i \) belonging to category \( c_1 \) or \( c_2 \) is classified into category \( \mu_i \) by the image classifier. As \( \mu_i \in \{c_1, c_2\} \), there are two probabilities corresponding to \( p(\mu_i|c_1) \): \( p(c_1|c_1) \) and \( p(c_2|c_1) \), and two probabilities corresponding to \( p(\mu_i|c_2) \): \( p(c_1|c_2) \) and \( p(c_2|c_2) \). The probability \( p(c_1|c_1) \) or \( p(c_2|c_1) \) represents the probability that an adult image is correctly classified as adult or mistakenly classified as non-adult, and the probability \( p(c_1|c_2) \) or \( p(c_2|c_2) \) represents the probability that a non-adult image is mistakenly classified as adult or correctly classified as non-adult. It is obvious that \( p(c_1|c_1) + p(c_2|c_1) = 1 \) and \( p(c_1|c_2) + p(c_2|c_2) = 1 \). The probability \( p(c_2|c_1) \) or \( p(c_2|c_2) \) can be estimated statistically by counting the numbers of images classified mistakenly by the image classifier in a large set of adult images or non-adult images.

An application of Bayes’ rule yields the following equations:

\[
p(c_1|R) = \frac{p(R|c_1)p(c_1)}{p(R)}, \tag{5}
\]

\[
p(c_2|R) = \frac{p(R|c_2)p(c_2)}{p(R)}. \tag{6}
\]

Equation (3) is substituted into (5), and (4) is substituted into (6). Then, the ratio \( f \) of the two posterior probabilities in (5) and (6) is represented as:

\[
f = \frac{p(c_1|R)}{p(c_2|R)} = \frac{p(R|c_1)p(c_1)}{p(R|c_2)p(c_2)} = \frac{\prod_{i=1}^{N} p(\mu_i|c_1)}{\prod_{i=1}^{N} p(\mu_i|c_2)} \times \frac{p(c_1)}{p(c_2)}. \tag{7}
\]

where \( p(c_1) \) or \( p(c_2) \) is the prior probability that the clip belongs to category \( c_1 \) or \( c_2 \).

The prior classification of an adult clip can be represented by prior classifications of all the images in the clip: \{\( p_1(c_1), p_2(c_1), \ldots, p_i(c_1), \ldots, p_N(c_1) \)\} and \{\( p_1(c_2), p_2(c_2), \ldots, p_i(c_2), \ldots, p_N(c_2) \)\}, where \( p_i(c_1) \) or \( p_i(c_2) \) is the probability that image \( i \) belongs to category \( c_1 \) or \( c_2 \)—a prior classification of image \( i \). Due to the independence between the classifications of the individual images, the following equations are obtained:

\[
p(c_1) = p_1(c_1)p_2(c_1) \cdots p_i(c_1) \cdots p_N(c_1), \tag{8}
\]

\[
p(c_2) = p_1(c_2)p_2(c_2) \cdots p_i(c_2) \cdots p_N(c_2). \tag{9}
\]

The substitution of Equations (8) and (9) into (7) yields the following equation:

\[
f = \frac{p(c_1|R)}{p(c_2|R)} = \frac{\prod_{i=1}^{N} (p(\mu_i|c_1) \times p_i(c_1))}{\prod_{i=1}^{N} (p(\mu_i|c_2) \times p_i(c_2))}. \tag{10}
\]

If \( f > 1 \), then the clip is classified as category \( c_1 \); otherwise as category \( c_2 \).

The remaining problem is to estimate \( p_i(c_1) \) or \( p_i(c_2) \) (It is obvious that \( p_i(c_1) + p_i(c_2) = 1 \)). As already mentioned, the audio accompanying a clip provides the information to obtain the prior classifications of the images in the clip. We use the audio section around image \( i \) to estimate \( p_i(c_1) \) or \( p_i(c_2) \): The prior probability \( p_i(c_1) \) or \( p_i(c_2) \) is replaced with the probability that the audio classifier classifies the audio section associated with image \( i \) as adult or non-adult.

For recognition of adult sounds, a number of adult sound samples are used to construct an adult sound model. The similarity between a test sound and the constructed adult sound model is used to estimate the prior probability that the image with which the sound is associated contains adult content. Recognition of adult sounds includes the extraction of audio features and the classification of the extracted audio features. Mel-Frequency Cepstral Coefficients (MFCC) [Sadka 2004; Buchanan 2005] are used as the sound features. The MFCC features for each audio frame are extracted by carrying out a discrete Fourier transform to obtain an amplitude spectrum, converting the amplitude spectrum into
a Mel-scale spectrum, and using the discrete cosine transform to convert a number of the logarithmic values of the Mel-weighted spectrum into 12 cepstral coefficients and the zeroth coefficient. A sound is represented by a vector of the 13 coefficients. The simple Gaussian mixture model (GMM) [Tax and Duin 2004] is used to estimate the distribution of the training set of adult sounds. As adult sounds are quite different from other sounds, the GMM, which only models the adult sounds, can accurately distinguish adult sounds from non-adult sounds. An appropriate value of the number of components in the GMM is estimated using the samples—the number of the components in the GMM is chosen such that the best performance is reached. The probability that a test sound matches the trained adult sound model is determined using the GMM probability distribution. If there is no available audio in a clip, the prior probability that each image belongs to the adult category is set to 0.5 since there is no prior knowledge.

To avoid the underflow produced by the product of the probability values, the log version of (10) is used to transform the product of the probability values to the sum of the logarithmic probability values:

\[
F = \log(f) = \sum_{i=1}^{N} (\log(p(\mu_i|c_1)) + \log(p_i(c_1)) - \log(p(\mu_i|c_2)) - \log(p_i(c_2))).
\] (11)

If \( F > 0 \), the clip is classified as adult.

4. RECOGNITION OF ADULT WEB PAGE BAGS

Features for a Web page are extracted according to the results of processing texts and recognizing images and videos in the Web page using the algorithms described in Sections 2 and 3. MIL is used to train a Web page classifier to identify adult Web pages from non-adult pages.

A Web page set, which consists of a Web page and its linked pages, is considered as adult if it has adult Web pages, and it is considered as non-adult if it has no adult Web page. Thus, we can select a predefined number of Web pages from a Web page set as instances to form a Web page bag, and then use MIL to recognize adult Web page bags. If a Web page bag is classified as adult, all the Web pages in the Web page set corresponding to this Web page bag is labeled as adult. Recognition of adult Web page bags using MIL can substantially increase the recognition efficiency in contrast to the methods that recognize Web pages one-by-one.

There are a number of MIL algorithms [Zhou et al. 2005], such as diverse density, Bayesian-kNN, citation-kNN [Wang and Zuchker 2000], multi-instance decision tree, neural networks, and support vector machines [Andrews et al. 2003]. In this article, four typical MIL algorithms, Bayesian-kNN, citation-kNN, mi-SVM, and MI-SVM, are applied to recognize adult Web page bags, because the models that they use are very simple, and they have low computational complexities and competitive accuracies [Wang and Zuchker 2000; Andrews et al. 2003].

4.1 Web Page Feature Extraction

4.1.1 Keyword construction and text feature extraction. The text feature is extracted based on keyword construction. We use the simple and effective document frequency [Yang 1997; Zhang and Zhou 2008] to construct keywords. The texts in a Web page form a document. The document frequency of a term is defined as the number of documents in which the term appears. In the training set, every term’s document frequency is counted. The terms with the highest document frequencies are selected as keywords, omitting empty words and stop words. The reason why we choose document frequency is that the estimation of document frequency does not use the Web pages’ label information, which is unknown for MIL.
Using the constructed set of keywords, the text feature for a Web page is extracted by assigning a weight to each keyword according to the number of times it appears in the Web page. We apply the term frequency inverse document frequency [Salton and Buckly 1998] to give a weight $w_i$ to a keyword $i$ for a Web page:

$$
 w_i = \frac{\Theta_i \cdot \log(E/e_i)}{\sqrt{\sum_{j=1}^{t} (\Theta_j \cdot \log(E/e_j))^2}},
$$

where $\Theta_i$ is the appearance frequency of keyword $i$ in the document, $E$ denotes the number of all the documents in the training set, $e_i$ denotes the document frequency of keyword $i$, and $t$ is the number of the keywords. The weight has the following characteristics.

— The more frequently a keyword appears in a document, the more important the keyword is for this document.
— The weight is inversely proportional to the document frequency. The fewer documents in which a keyword appears, the more classification information it provides for classifying documents.
— The weight considers the keyword appearance frequency and document frequency using the product operation as shown in (12). The product is normalized by the denominator. The normalization removes the impact of document length on the weight. A keyword with a low frequency in a document can obtain a high weight if the length of the document is short.

4.1.2 Image and Video Features. We extract the following image and video features for an instance according to the recognition results for the images and videos in the corresponding Web page.

— the ratio of the number of the large images which are recognized as adult to the total number of large images where large images are defined as ones containing more than 50,000 pixels;
— the ratio of the number of the medium-sized images which are recognized as adult to the total number of medium-sized images where medium-sized images are defined as ones containing 10,000 to 50,000 pixels;
— the ratio of the number of the small images which are recognized as adult to the total number of small images where small images are defined as ones containing less than 10,000 pixels;
— the weighted sum of regression outputs of the random forest-based classifier for all the images in the Web page where the regression response of the classifier is the average of the responses over all the decision trees in the forest, and the weights are obtained by normalizing the numbers of pixels in the images.
— the ratio of the number of the long videos which are recognized as adult to the total number of long videos where long videos are defined as ones with duration of more than 3.5 minutes;
— the ratio of the number of the medium-sized videos which are recognized as adult to the total number of medium-sized videos where medium-sized videos are defined as ones with duration of 1 to 3.5 minutes;
— the ratio of the number of the short videos which are recognized as adult to the total number of short videos where short videos are defined as ones with duration of less than 1 minute.

Image features are extracted according to image sizes as different image sizes have different contributions to Web page content. Video features are extracted according to video durations, as longer videos more easily mistakenly classified. The preceding seven features are normalized into a 7-dimensional vector.
4.1.3 Similarity estimation. Each Web page is represented by a vector \( W = (w_1, w_2, \ldots, w_t, w_{t+1}, \ldots, w_{t+7}) \), where \( w_i \) (\( 1 \leq i \leq t \)) represents the weight of the \( i \)-th keyword in the document and \( w_{t+j} \) (\( 1 \leq j \leq 7 \)) represents the \( j \)-th image/video feature of the Web page. The cosine similarity is used to measure the similarity between the feature vectors of any two Web pages. For two feature vectors \( W_q \) and \( W_d \), their cosine similarity \( \cos(W_q, W_d) \) is described as:

\[
\cos(W_q, W_d) = \frac{\sum_{i=1}^{\vartheta} w_{qi} w_{di}}{\sqrt{\sum_{i=1}^{\vartheta} w_{qi}^2 \cdot \sum_{i=1}^{\vartheta} w_{di}^2}},
\]

where \( \vartheta = t + 7 \).

4.2 Multi-Instance Learning

In the following, we briefly describe the methods for recognizing adult Web pages using Bayesian-kNN, citation-kNN, mi-SVM, and MI-SVM.

4.2.1 Bayesian-kNN and citation-kNN. Bayesian-kNN and citation-kNN are derived from kNN. The similarities between Web page bags are measured using the minimum Hausdorff distance which is defined as:

\[
h(A, B) = \min_{a \in A} \min_{b \in B} ||a - b|| = h(B, A),
\]

where \( a \) is a Web page instance in bag \( A \), and \( b \) is a Web page instance in bag \( B \). It has been verified in Wang and Zuckerman [2000] that the minimum Hausdorff distance is more effective than the standard maximum Hausdorff distance for estimating the similarities between bags.

4.2.1.1 Bayesian-kNN. For a test Web page bag \( A \), we find \( k \) sample Web page bags \( \{B_1, B_2, \ldots, B_k\} \) whose minimum Hausdorff distances with bag \( A \) are minimum from all the sample Web page bags. Their labels are denoted using \( \{\mu_1, \mu_2, \ldots, \mu_k\} \), respectively, where \( \mu_i \) belongs to either the adult category \( c_1 \) or the non-adult category \( c_2 \). According to Bayes’ rule, the label \( c \ (c \in \{c_1, c_2\}) \) of bag \( A \) is determined by:

\[
c = \arg \max_{\mu \in \{c_1, c_2\}} p(\mu | \mu_1, \mu_2, \ldots, \mu_k)
= \arg \max_{\mu \in \{c_1, c_2\}} \frac{p(\mu_1, \mu_2, \ldots, \mu_k | \mu)p(\mu)}{p(\mu_1, \mu_2, \ldots, \mu_k)}
= \arg \max_{\mu \in \{c_1, c_2\}} p(\mu_1, \mu_2, \ldots, \mu_k | \mu)p(\mu)
\]

The probabilities \( p(\mu_1, \mu_2, \ldots, \mu_k | \mu) \) and \( p(\mu) \) are estimated from the training set. For each Web page bag sample, its \( k \) nearest Web page bag samples are found. We count the number \( g \) of the samples whose nearest samples have the same number of samples with label \( \mu \) as the number of label \( \mu \) in \( p(\mu_1, \mu_2, \ldots, \mu_k | \mu) \). Then, \( p(\mu_1, \mu_2, \ldots, \mu_k | \mu) \) is the ratio of \( g \) to the number of all the samples. The probability \( p(\mu) \) is estimated as the proportion of the samples with label \( \mu \) in all the Web page bag samples.

4.2.1.2 Citation-kNN. The citation-kNN algorithm considers not only the labels of the Web page bag samples whose feature vectors are nearest to the feature vector of the test Web page bag, but also the labels of the Web page bag samples whose nearest samples contain the test sample. We find \( r \) Web page bag samples (references) nearest to the test Web page bag sample and \( c \) Web page bag samples (citers) whose nearest Web page bag samples contain the test Web page bag sample. As citers are more important than references owing to the effect of pseudopositive instances, \( c \) is selected to be larger.
than $r$. Let $r_a$ and $r_n$ be the numbers of adult and non-adult references, $r = r_a + r_n$. Let $c_a$ and $c_n$ be the numbers of adult and non-adult citers, $c = c_a + c_n$. Let $\pi_a$ be the sum of the number of the adult references and the number of the adult citers: $\pi_a = r_a + c_a$. Let $\pi_n$ be the sum of the number of the non-adult references and the number of the non-adult citers: $\pi_n = r_n + c_n$. If $\pi_a > \pi_n$, the test Web page bag is labeled as adult; otherwise it is labeled as non-adult.

4.2.2 mi-SVM and MI-SVM. The mi-SVM and MI-SVM Andrews et al. [2003] are extended from SVM to deal with MIL problems.

The mi-SVM looks for the hyper-plane such that for each adult Web page bag there is at least one Web page lying in the positive half-space, and all the Web pages belonging to non-adult Web page bags lie in the negative half-space. The labels of the Web pages in the sample bags are initially set to the labels of their bags, and the Web pages are classified using SVM. In each of the adult bags which are misclassified as non-adult, its Web page, which lies nearest to the hyper-plane, is labeled as adult, and the other Web pages are set to the labels to which they are classified by the SVM. Then all the Web pages with new labels are classified again using SVM. The preceding process is repeated until the hyper-plane has little change.

The MI-SVM represents an adult bag using the instance in which it is farthest from the hyper-plane. Each adult bag is initially represented by the average vector of all the Web pages in the bag, and the average vector is labeled as adult. The Web pages represented by these average vectors, together with the Web pages in the non-adult bags, are classified using SVM. Each of the adult bags is represented by its Web page lying farthest from the hyper-plane. The Web pages representing the adult bags are labeled as adult. They, together with the Web pages in the non-adult bags, are classified again using SVM. The process is repeated until the hyper-plane has little change.

5. EXPERIMENTAL RESULTS

All of the preceding algorithms and the system were implemented using Microsoft Visual C++ 2005 on the Windows Vista platform. The runtimes are measured on a P4-3.2G computer. In the following, we evaluate in succession the performance of the recognition of adult images, the recognition of adult videos, and the recognition of adult Web page bags.

5.1 Skin Detection and Adult Image Recognition

5.1.1 Skin detection. We collected 17,812 skin patches and 31,623 non-skin patches. The features extracted from these patches were used to train a random forest-based classifier whose number of decision trees was set to 1000.

Figure 2 shows examples of comparisons of skin detection results between our algorithm and other classical algorithms where the first column corresponds to the original images, the second to the results of the threshold boundary model-based algorithm in Kovac et al. [2003], the third to the results of the ellipse model-based algorithm in Hsu et al. [2002], the fourth to the results of the Gaussian model-based algorithm in Lee and Yoo [2002], the fifth to the results of the GMM-based algorithm in Jones and Rehg [2002], the sixth to the results of our regular patch-based algorithm, the seventh to the results of our irregular patch-based algorithm, and the eighth to the results of our irregular patch-based algorithm. The results of our algorithms are overall more accurate than the results of the competing algorithms. Our irregular patch-based algorithm obtains the most satisfactory skin detection results, but it is slow. Our regular patch-based algorithm is faster, but the obtained skin region boundaries are jagged.

To further compare our algorithm with other skin detection algorithms, we tested our algorithm using the Compaq skin image database [Jones and Rehg 2002], which has been widely used for skin
Fig. 2. Examples of comparisons of skin detection results between our algorithms and other classical algorithms: (a) original images; (b) Kovac et al.’s algorithm [2003]; (c) Hsu et al.’s algorithm [2003]; (d) the Gauss-based algorithm; (e) the GMM-based algorithm; (f) our regular patch-based algorithm; (g) segmentation using Felzenszwalb and Huttenlocher’s algorithm [2004]; (h) our irregular patch-based algorithm.

Table I. Comparison in the Compaq Skin Image Database

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Detection rate</th>
<th>False alarm rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian model [Lee and Yoo 2002]</td>
<td>90%</td>
<td>33%</td>
</tr>
<tr>
<td>GMM [Jones and Kehg 2002]</td>
<td>90%</td>
<td>15.5%</td>
</tr>
<tr>
<td>Ellipse model [Lee and Yoo 2002]</td>
<td>90%</td>
<td>20.9%</td>
</tr>
<tr>
<td>Maximum entropy [Jedynak 2002]</td>
<td>86%</td>
<td>8%</td>
</tr>
<tr>
<td>SOM [Brown et al. 2001]</td>
<td>78%</td>
<td>32%</td>
</tr>
<tr>
<td>Our regular patch-based</td>
<td>93.8%</td>
<td>11.3%</td>
</tr>
<tr>
<td>Our irregular patch-based</td>
<td>94.1%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>

detection. Table I shows the comparison of the detection rate and the false alarm rate between our algorithms and other algorithms, where detection rates and false alarm rates of other algorithms were derived from the corresponding published papers. From the table, it is seen that our algorithms obtain the highest detection rate, while maintaining comparatively low false alarm rates. Furthermore, our regular patch-based algorithm is very fast due to the use of integral images, while obtaining more accurate skin detection results than other approaches. The reason why the best results were obtained by our algorithms is that we directly modeled skin patches by representing relations between values of pixels in a patch, and directly classifying test patches as skin or non-skin. However, the competing algorithms classify each pixel as skin or non-skin individually, considering or not considering the relations between pixel values in a patch. Consequently, the global skin information in a patch is more accurately modeled in our algorithm.

5.1.2 Adult image recognition. Our training set consisted of 13,649 images, 3,784 of which were adult and 9,865 of which were non-adult. The test sets consisted of a set of low-resolution images and a set of high-resolution images. Sizes of low-resolution images were usually lower than 300*300. Sizes of high-resolution images were usually higher than 800*800. The low-resolution set was further divided into three subsets: one consisting of 200 adult images and 200 non-adult images, which are mainly Bikini images; one consisting of 500 adult images and 500 non-adult images, which are mainly animal images; and one consisting of 1000 adult images and 1000 non-adult images, which are mainly...
Table II. Comparison between Adult Images Recognition Results of Different Algorithms

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
<td>F1</td>
</tr>
<tr>
<td>Low-resolution set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subset 1</td>
<td>87.5%</td>
<td>77.8%</td>
<td>82.4%</td>
</tr>
<tr>
<td>Subset 2</td>
<td>67.7%</td>
<td>94.4%</td>
<td>78.9%</td>
</tr>
<tr>
<td>Subset 3</td>
<td>76.4%</td>
<td>84.7%</td>
<td>84.7%</td>
</tr>
<tr>
<td>High-resolution set</td>
<td>98.9%</td>
<td>86.0%</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

face images and natural landscape images. The high-resolution set consisted of 845 adult images and 1108 non-adult images.

Table II compares the results of our algorithm with the results of two state-of-the-art representative and typical adult-image recognition algorithms: Zheng et al. [2004] (the poesia algorithm) and Liang et al. [2004]. In the table, recall is the fraction of the correctly recognized adult images in the set of the adult images; the precision is the fraction of the correctly recognized adult images in the set of images which are recognized as adult for all the images in the test set; and the measure $F_1$, which is a performance criterion derived from precision and recall, is defined as:

$$F_1 = \frac{\text{Recall} \times \text{Precision} \times 2}{\text{Recall} + \text{Precision}}.$$  \hspace{1cm} (16)

The higher the recall, the precision, and the measure $F_1$, the better the performance of an algorithm. From the table, it is shown that our algorithm is faster than the competing algorithms and the recognition results of our algorithm are overall more accurate than those of the competing algorithms. Especially, for high-resolution images, our algorithm obtains much more accurate results. The reasons why our algorithm is superior are that more accurate skin detection is obtained and more effective and comprehensive features are used.

5.2 Recognition of Adult Videos

In the following, the results for recognizing adult sounds, adult video frames, and adult videos are described.

5.2.1 Recognition of adult sounds. In the experiments, the audio waveform was transformed into a sequence of 13-dimensional feature vectors. For simplicity and reduction of computational cost, each sound is represented as the mean of the MFCC vectors in the sequence.

To evaluate the performance of our GMM-based algorithm for recognizing adult sounds, 1,412 sounds were collected of which 592 adult sounds were used as the training data and 820 sounds (including 268 adult sounds and 552 non-adult sounds) were used as the test data. The topics of the non-adult sounds include animals, birds, nature, musical instruments, speech, vehicles, sports, and recreation. The sounds vary in duration from less than a second to about 30 seconds. All audio streams are in the 22,050Hz, 16-bit and mono-channel format. They are divided into frames of 16ms with 50% overlap for feature extraction.

Table III shows the performance of the GMM-based algorithm for recognizing adult sounds with different numbers of mixture components. From the table, it is shown that recall, precision, and the measure $F_1$ overall increase, with minor fluctuation when the number of the Gaussian mixture components increases from 1 to 15. When the number of the components is more than 15, the performance of the model reaches saturation and begins to degrade due to the overfitting. So, when the number of the mixture components is set to 15, the algorithm reaches its best performance.

5.2.2 Recognition of adult video frames. The video frames used were extracted from a set of videos. They consisted of 53,554 adult frames and 83,548 non-adult frames. Table IV shows the results of
Table III. The Results of Recognizing Adult Sounds with Different Numbers of Mixture Components

<table>
<thead>
<tr>
<th>Number of components</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>96.9%</td>
<td>87.6%</td>
<td>92.0%</td>
</tr>
<tr>
<td>12</td>
<td>96.9%</td>
<td>86.4%</td>
<td>91.3%</td>
</tr>
<tr>
<td>13</td>
<td>96.4%</td>
<td>84.6%</td>
<td>90.1%</td>
</tr>
<tr>
<td>14</td>
<td>96.4%</td>
<td>91.5%</td>
<td>93.9%</td>
</tr>
<tr>
<td>15</td>
<td>97.2%</td>
<td>91.4%</td>
<td>94.2%</td>
</tr>
<tr>
<td>16</td>
<td>96.4%</td>
<td>88.3%</td>
<td>92.2%</td>
</tr>
<tr>
<td>17</td>
<td>95.7%</td>
<td>87.8%</td>
<td>91.6%</td>
</tr>
<tr>
<td>18</td>
<td>98.0%</td>
<td>87.5%</td>
<td>92.4%</td>
</tr>
<tr>
<td>19</td>
<td>96.9%</td>
<td>90.9%</td>
<td>93.8%</td>
</tr>
<tr>
<td>20</td>
<td>96.2%</td>
<td>84.5%</td>
<td>90.0%</td>
</tr>
</tbody>
</table>

Table IV. The Results of Recognizing Adult Video Frames

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.4%</td>
<td>79.7%</td>
<td>83.4%</td>
</tr>
</tbody>
</table>

Our adult-image recognition algorithm for these frames. The error rate for the frames from videos is much higher than that for high-resolution Web images shown in Table II and as much as those for low-resolution Web images. This is due to the low image quality and the higher complexity of scenes in videos.

5.2.3 Recognition of adult videos. We apply the following five classifiers to recognize adult videos:

1. The image-based classifier derived from [Wang et al. 1998]. The number of detected adult frames is counted. If a sufficient number of adult frames are detected, the video is labeled as adult.
2. The sound-based classifier. Only audio information is used to recognize adult videos.
3. The early fusion strategy. The normalized feature vector for an image and the normalized feature vector for its associated audio section are concatenated to form a feature vector which is classified by a random forest-based classifier. If a sufficient number of images and their associated audio sections in a video are classified as adult, then this video is classified as adult.
4. The classifier extended from our previous work. The approach in Hu et al. [2007] for recognizing adult image Web pages is extended to recognize adult videos by using the probability that the whole sound in a video is recognized as adult as the prior probability that the video is adult [Zuo et al. 2008].
5. The proposed classifier. The probability that the sound section associated with a frame is recognized as adult is used as the prior probability that the frame is adult.

Both the last two classifiers adopt the late fusion strategy.

The frames extracted from the training videos were used to estimate the probabilities $p(c_1|c_1)$, $p(c_2|c_1)$, $p(c_1|c_2)$, and $p(c_2|c_2)$ described in Section 3, where category $c_1$ corresponds to the adult video category and category $c_2$ corresponds to the non-adult video category. These probabilities are estimated as: $p(c_1|c_1) = 0.874$, $p(c_2|c_1) = 0.126$, $p(c_1|c_2) = 0.142$, and $p(c_2|c_2) = 0.858$.

We downloaded 352 adult videos and 537 non-adult videos from the Internet to evaluate the performance of the five classifiers. The average and total numbers of frames in the adult videos were 4,680 and 1,647,361, respectively. Those in the non-adult videos were 3,863 and 3,074,432, respectively.
Table V. The Results of Different Classifiers for Recognizing Adult Videos

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image-based classifier [Wang, 1998]</td>
<td>85.2%</td>
<td>81.1%</td>
<td>83.1%</td>
</tr>
<tr>
<td>Sound-based classifier</td>
<td>90.3%</td>
<td>83.0%</td>
<td>86.5%</td>
</tr>
<tr>
<td>Early fusion strategy</td>
<td>94.9%</td>
<td>88.6%</td>
<td>91.6%</td>
</tr>
<tr>
<td>Classifier extended from [Hu, 2007]</td>
<td>98.3%</td>
<td>92.5%</td>
<td>95.3%</td>
</tr>
<tr>
<td>Classifier in this paper</td>
<td>99.1%</td>
<td>93.8%</td>
<td>96.4%</td>
</tr>
</tbody>
</table>

sound section associated with each frame is 1 second long, 1/2 second before the frame and 1/2 second following the frame.

Table V shows the adult video recognition results of the image-based classifier, the sound-based classifier, the earlier fusion strategy, the classifier extended from Hu et al. [2007], and the classifier in this article. From the table, the following useful points are noted.

—The results of the sound-based classifier which uses audio information alone are better than the results of the image-based classifier which uses image information alone. This partly illustrates the distinguishability of the adult sounds from other sounds and the reasonableness of using a GMM-based classifier, which does not model non-adult sounds to recognize adult sounds.

—The results of the early fusion strategy are more accurate than the results of both the image-based classifier and the audio-based classifier.

—Both the classifier extended from Hu et al. [2007] and the classifier in this article, obtain more accurate results than the early fusion strategy.

—The performance of the classifier in this article is superior to the performance of the classifier extended from Hu et al. [2007].

The reason why the early fusion strategy obtains less accurate results than the late fusion strategy is that the early fusion strategy achieves the fusion by simply concatenating the image features and the audio features, and the probability relations between images and audio sections are not represented. The reason why the classifier in this article outperforms the classifier extended from Hu et al. [2007] is that more information from the audio is used in the classifier in this article. In the extended classifier, only one probability, produced by recognizing the whole audio, contributes to the adult video recognition. In the classifier in this article, many probabilities produced by recognizing the audio sections associated with many frames contribute to the recognition of the video, and then the temporal correspondence between the audio sequence and the video frame sequence is used in the fusion process. Thus, the classifier in this article is more accurate at recognizing adult videos than the extended classifier.

5.3 Recognition of Adult Web Page Bags

We collected 3,090 Web page bags from the Internet where 1,957 adult bags were collected from many adult websites, and 1,133 non-adult bags were collected from many Web sites that cover a variety of subjects, such as news, education, finance, health, military, information technology, travel, and shopping. Some pages were omitted from some bags in order to ensure that there are no overlapped pages between any two bags. We randomly selected 198 bags (which included 141 adult bags and 57 non-adult bags) as the training set, while the other bags were used as the test set.

Table VI shows the average values of recall, precision, and F1 of the four MIL algorithms, Bayesian-kNN, citation-kNN, mi-SVM, and MI-SVM with different numbers of web pages selected per bag. The


Table VI. The Results (%) of Bayesian-kNN, Citation-kNN, mi-SVM, and MI-SVM

<table>
<thead>
<tr>
<th>Number of instances per bag</th>
<th>Bayesian-kNN</th>
<th>Citation-kNN</th>
<th>mi-SVM</th>
<th>MI-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
<td>F1</td>
<td>Recall</td>
</tr>
<tr>
<td>5</td>
<td>96.3</td>
<td>96.8</td>
<td>96.6</td>
<td>96.4</td>
</tr>
<tr>
<td>6</td>
<td>97.1</td>
<td>96.8</td>
<td>97.0</td>
<td>98.4</td>
</tr>
<tr>
<td>7</td>
<td>96.7</td>
<td>97.4</td>
<td>97.0</td>
<td>98.4</td>
</tr>
<tr>
<td>8</td>
<td>97.7</td>
<td>96.8</td>
<td>97.2</td>
<td>98.7</td>
</tr>
<tr>
<td>9</td>
<td>97.5</td>
<td>97.2</td>
<td>97.4</td>
<td>98.5</td>
</tr>
<tr>
<td>10</td>
<td>97.9</td>
<td>97.1</td>
<td>97.5</td>
<td>98.5</td>
</tr>
</tbody>
</table>

Table VII. Comparison Between MIL-Based Algorithms and the Algorithms Recognizing Web Pages One-By-One

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall %</th>
<th>Precision %</th>
<th>F1 %</th>
<th>Time per page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The text-based algorithm [Du et al. 2003]</td>
<td>83.1%</td>
<td>81.8%</td>
<td>82.4%</td>
<td>0.61s</td>
</tr>
<tr>
<td>The image-based algorithm [Wang et al. 1998]</td>
<td>80.3%</td>
<td>79.9%</td>
<td>80.1%</td>
<td>1.38s</td>
</tr>
<tr>
<td>The text and image combined algorithm [Jones and Rehg 2002]</td>
<td>90.1%</td>
<td>80.3%</td>
<td>84.9%</td>
<td>1.23s</td>
</tr>
<tr>
<td>kNN: one by one</td>
<td>86.6%</td>
<td>84.8%</td>
<td>85.7%</td>
<td>1.16s</td>
</tr>
<tr>
<td>Bayesian-kNN</td>
<td>97.5%</td>
<td>96.7%</td>
<td>97.1%</td>
<td>0.11s</td>
</tr>
<tr>
<td>Citation-kNN</td>
<td>98.3%</td>
<td>96.9%</td>
<td>97.6%</td>
<td>0.13s</td>
</tr>
<tr>
<td>mi-SVM</td>
<td>99.0%</td>
<td>96.8%</td>
<td>97.9%</td>
<td>0.06s</td>
</tr>
<tr>
<td>MI-SVM</td>
<td>98.8%</td>
<td>97.7%</td>
<td>98.2%</td>
<td>0.07s</td>
</tr>
</tbody>
</table>

average values are calculated from the data obtained by taking the number of keywords from 50, 100, 200, 300, 400, and 500. From the experiments, the following useful points are noted.

—All the Bayesian-kNN, citation-kNN, mi-SVM, and MI-SVM perform effectively in the recognition of adult Web page bags.
—Citation-kNN obtains more accurate results than Bayesian-kNN.
—The results of mi-SVM are slightly more accurate than the results of citation-kNN.
—The results of MI-SVM are slightly more accurate than the results of mi-SVM.
—There is little performance difference between 5, 6, 7, 8, 9, and 10 Web pages per bag. This is because in adult Web sites, almost every Web page is adult. Therefore, the “intuitive” claim that the more the Web pages selected per bag, the better the result is does not hold true.

The reason why the results of mi-SVM and MI-SVM are more accurate than the results of Bayesian-kNN and citation-kNN is that SVM has more powerful classification ability than kNN.

We compared our MIL-based adult Web page recognition algorithms with three of the state of the art adult Web page recognition algorithms: the text-based algorithm in Du et al. [2003], the image-based algorithm in Wang et al. [1998], and the text and image combined algorithm in Jones and Rehg [2002]. We also compared the MIL-based algorithms with the kNN algorithm whose parameters are set in the same way as Bayesian-kNN and citation-kNN do. The competing algorithms classify Web pages one-by-one. The comparison results are shown in Table VII. From the table, it is shown that the MIL-based algorithms not only obtain more accurate results, but also require much less runtime than the competing algorithms. The reason why our MIL-based algorithms obtain more accurate results than the text-based algorithm in Du et al. [2003] and the image-based algorithm in Wang et al. [1998] is that both text information and image information are effectively combined in our algorithm. The reason why the text and image combined algorithm in Jones and Rehg [2002] is inferior to our algorithms is that their algorithm classifies a Web page as adult if either the text classifier or the image classifier finds adult content in it. The result is an increased false acceptance rate. The reason why the MIL-based algorithms need much less runtime than the algorithms recognizing Web pages one-by-one is...
that only a small number of Web pages are selected from a Web page set which consists of a Web page and its linked pages to form a bag, and all the web pages in the Web page set are labeled according to the classification result of the bag.

6. CONCLUSION

In this article, we have developed a system which can recognize Web adult images, adult videos, and adult Web page bags. In our adult-image recognition algorithm, skin patches have been detected by modeling skin patches rather than skin pixels, and the recognition features have been extracted using the idea of going from global to local. In our algorithm for recognizing adult videos, the result of recognizing the audio section associated with an image in a video has been used as a prior classification of the image. Our algorithm has achieved a better performance than the ones which use image information alone or audio information alone. We have carried out recognition of adult Web page bags rather than individual Web pages. Bayesian-kNN citation-kNN, mi-SVM, and MI-SVM have been used to recognize adult Web page bags, respectively. The results are very promising.

Our future work will focus on the following areas.

—We will use asymmetric bagging and random subspace for SVM-based image retrieval [Tao et al. 2006], geometric mean-based subspace selection [Tao et al. 2009], and max-min distance for dimension reduction [Bian and Tao 2011] to select more effective features and design more effective classification methods to recognize adult videos and adult Web pages.

—We will extend our algorithms to recognize other types of socially sensitive information, such as drug abuse and violence.

APPENDIX

Hierarchical Handling of Web Adult Images

A rapid algorithm is proposed to calculate the probability of human existence in an image, and images without human existence are filtered out. In the following, we present our algorithm for estimating the probability of human existence in an image. The estimation of probabilities from an image, its associated text, and the corresponding Web page title, the fusion of these probabilities, as well as the corresponding experimental results are described.

A.1 Probability from Image

The skin detection result for an image is used to compute the human existence probability \( p_{\text{image}} \) from the image. The image is equally divided into \( m_1 \times m_2 \) blocks [Wu et al. 2008] (in our experiments, both \( m_1 \) and \( m_2 \) are chosen as 8 for an empirical balance between the accuracy and the complexity). If the proportion of skin pixels in a block exceeds a predefined threshold \( \tau_{\text{skin}} \), then the block is taken as a skin block; otherwise, it is a non-skin block. Two blocks are considered as connected only if they share a side. Connected blocks are combined into a skin area. For an image, we construct a set \( \Omega \) which consists of the numbers of skin blocks in each skin area: \( \Omega = \{n_1, \ldots, n_j, \ldots, n_J\} \), where \( J \) is the number of the detected skin areas in an image and \( n_j \) is the number of the skin blocks in the \( j \)th skin area. Figure 3 shows an example of the process of generating the set \( \Omega \), where (a), (b), and (c) show, respectively, the original image, the detected skin pixels, and the skin areas. In this example, there are 13 skin blocks which make up 5 skin areas. Its \( \Omega \) set is \( \{5, 4, 2, 1, 1\} \). The \( \Omega \) set reflects the statistical information about the distribution of skin pixels in an image. The problem is then reduced to that of determining human existence in an image according to the \( \Omega \) set of the image.

Considering that a non-skin block may be misclassified as a skin block, we introduce the probability \( p_{\text{error}} \) that a block without human skin is misclassified as a skin block. The probability \( p_{\text{error}} \) depends...
on the factors of the skin detector, the block size, and the threshold $\tau_{\text{skin}}$. If all these factors are fixed, $p_{\text{error}}$ is fixed and it can be considered as a constant.

Let $\xi_j$ represent the event that the $j$th skin area in an image is mistakenly detected, that is, in fact this area is not a skin area. Let $\Omega_j$ be the $j$th element in the $\Omega$ set of the image. Then,

$$p(\Omega_j|\xi_j) = (p_{\text{error}})^{nj}. \quad (A)$$

According to Bayes’ rule, the following equation is obtained:

$$p(\xi_j|\Omega_j) = \frac{p(\Omega_j|\xi_j) \times p(\xi_j)}{p(\Omega_j)} = \frac{(p_{\text{error}})^{nj} \times p(\xi_j)}{p(\Omega_j)}. \quad (B)$$

The probability $p(\xi_j|\Omega_j)$ in (B) denotes the probability that there is no skin in the $j$th area on the condition that the $j$th skin area is detected. It is noted that if there is no human skin in the image, none of the detected skin areas is from human skin. Let $\not\Omega$ be the event that there is human skin in the detected skin areas, that is, not all the skin areas in the image are detected mistakenly. If $j \neq i$, $\xi_j$ is independent of $\Omega_i$. Then, the Equation (C) is obtained:

$$p(\not\Omega|\Omega_j) = 1 - p(\{\xi_j\}|\Omega_j)$$
$$= 1 - \prod_j p(\xi_j|\Omega_j)$$
$$= 1 - (p_{\text{error}})^N \prod_j \frac{p(\xi_j)}{p(\Omega_j)}. \quad (C)$$

where $N$ is the total number of the detected skin blocks in the image, that is, $N = \sum_j n_j$. We estimate human existence in an image using $p(\not\Omega|\Omega)$ approximately:

$$p_{\text{image}} \approx p(\not\Omega|\Omega). \quad (D)$$

The probability $p(\xi_j)$ is independent of the appointed area. Then, $p(\xi_j)$ can be considered as a constant represented as $p_{\text{mis}}$, which is further defined as the prior probability that a skin area is mistakenly detected, and (D) is transformed to:

$$p_{\text{image}} \approx 1 - (p_{\text{error}})^N \times (p_{\text{mis}})^J. \quad (E)$$

Let $p_{\text{tos}}$ be the probability that a random block is detected as a skin block. The prior probability $p(\Omega_j)$ is obtained by $p(\Omega_j) = (p_{\text{tos}})^{nj}$, as $p_{\text{tos}}$ is defined for a random block. Then, (E) is transformed to:

$$p_{\text{image}} \approx 1 - \left(\frac{p_{\text{error}}}{p_{\text{tos}}}\right)^N \times (p_{\text{mis}})^J. \quad (F)$$

Fig. 3. An example of generating the $\Omega$ set: (a) original image; (b) skin pixels; (c) skin blocks and areas.
The parameters $p_{\text{error}}$, $p_{\text{mis}}$, and $p_{\text{tos}}$ are learned from a large-size training image set. Given a large image collection $\{\mathbb{N}_1, \mathbb{N}_2, \ldots, \mathbb{N}_\xi\}$ where $\xi$ is the number of images in the set, each image $\mathbb{N}_i$ is labeled manually and its label is represented as $y_i$. If it is labeled as human existence, then $y_i$ is set to 1; otherwise it is set to 0. The numbers of the detected skin blocks and skin areas in image $\mathbb{N}_i$ are denoted as $N_i$ and $J_i$. The parameters are estimated to minimize the following error function, which is defined according to (F), in order that images with human existence have as large $p_{\text{image}}$ as possible and images without human existence have as less $p_{\text{image}}$ as possible:

$$\min_{i=1}^{\xi} \left( 1 - \left( \frac{p_{\text{error}}}{p_{\text{tos}}} \right)^{N_i} \left( p_{\text{mis}} \right)^{J_i} - y_i \right)^2.$$  

(G)

First, each parameter is sampled from ten values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. The (approximately) optimal parameter combination is performed by exhaustive search in the space $[p_{\text{error}}, p_{\text{mis}}, p_{\text{tos}}]$. Then, the obtained parameters are further sampled. For example, if a parameter is calculated as 0.6, then it is further sampled from 0.55, 0.56, 0.57, 0.58, 0.59, 0.60, 0.61, 0.62, 0.63, 0.64, 0.65. The combination of the three parameters is further optimized by searching on the sampled values. Based on the estimated values of these parameters, the human existence probability $p_{\text{image}}$ for a test image is determined using (F).

A.2 Probabilities from Associated Text and Title

The text associated with an image in a Web page is usually related to the semantic description of the image [Gao et al. 2005]. Thus, the associated text is useful to determine human existence in the image. The result of classifying the text reflects the probability $p_{\text{text}}$ of the human existence in the image based on the text. A Web page title usually summarizes the theme of the Web page. Titles are useful for Web page classification [Yu et al. 2004], and a Web page title contains the clues of human existence of images in the Web page. We can obtain a probability $p_{\text{title}}$ of human existence in an image from the title of the corresponding Web page.

We train two naive Bayesian classifiers for texts associated with images and Web page titles. The reasons why the naive Bayesian classifier, which assumes that the feature vector components are independent of each other, is used are as follows.

—The semantic associations in texts associated with images can be ignored and terms in such texts are relatively independent.
—The naive Bayesian classifier outputs the probability of human existence rather than makes a yes/no decision on human existence, and this probability is used in the probability fusion in the later stages.

Text features are represented by each keyword’s term frequency inverse document frequency as described in Section 4.1.1. Let $C = \{c_1, c_2\}$ be the set of categories where $c_1$ and $c_2$ represent the human existence category and the non-human existence category, respectively. Given a new feature vector $w$, the naive Bayesian classifier is used to calculate the probability $p(c_j|w)$ ($c_j \in \{c_1, c_2\}$). The probability $p(c_1|w)$ from the classifier for texts associated with images is used as $p_{\text{text}}$. The probability $p(c_1|w)$ from the classifier for titles is used as $p_{\text{title}}$.

Three things should be noted. (1) During the training stage, the text associated with each image is labeled as human existence or non-human existence according to the category of the image. The title of a Web page is labeled as human existence if in more than half of the images in the Web page humans exist. (2) Both the texts associated with images and the Web page titles are short—each of them usually contains no more than twenty words. Consequently, all the words except the empty words and stop words are selected as keywords and all the people’s names are taken as the same feature,
as different names provide the same information for determining human existence. (3) If there is no associated text or title for an image, the corresponding probability $p_{text}$ or $p_{title}$ is set to 0.5.

A.3 Data Fusion

After the three probabilities $p_{image}$, $p_{text}$, and $p_{title}$ for an image are obtained, the final probability $p_{human}$ of human existence in the image is obtained by fusing them. We apply the Dempster-Shafer rule [Zhu and Basir 2003] to combine the three probabilities $p_{image}$, $p_{text}$, and $p_{title}$. The equations for obtaining probability $p_{human}$ are given as follows, where $\gamma_1$, $\gamma_2$, $v_1$, and $v_2$ are temporary variables introduced for maximizing computational efficiency.

$$\gamma_1 = p_{text}(1 - p_{text}) + p_{title}(1 - p_{title}) \quad \text{(H)}$$

$$v_1 = \frac{1}{1 - \gamma_1} p_{text} \cdot p_{title} \quad \text{(I)}$$

$$v_2 = \frac{1}{1 - \gamma_1} (1 - p_{text})(1 - p_{title}) \quad \text{(J)}$$

$$\gamma_2 = v_1 \cdot (1 - p_{image}) + v_2 \cdot p_{image} \quad \text{(K)}$$

$$p_{human} = \frac{1}{1 - \gamma_2} v_1 \cdot p_{image}. \quad \text{(L)}$$

If $p_{human}$ is larger than a threshold $\tau_{prob}$, then it is determined that there is human existence in the image and the adult-image recognition algorithm is used to determine whether the image contains adult content.

A.4 Experiments

We evaluated whether our hierarchical handling method can improve the performance of Web adult-image recognition by comparing our method with the method using the adult-image recognition algorithm alone. We collected 1,610 images with the associated texts and titles from 1,000 Web pages and used them as the training set. In these images, there were 829 images with human existence and 781 images without human existence, where the images with human existence consisted of 390 non-adult images and 439 adult images. These 1,000 Web pages cover 10 topics: arts, business, science, computer, news, shopping, game/recreation, society, health, and sports. We collected 832 images with the associated texts and titles from the Internet and used them as the test set, which includes 421 non-adult images and 411 adult images.

The parameters of the two naive Bayesian classifiers for texts associated with images and titles were learned from the 1,610 training samples. The parameters $\tau_{skin}$, $p_{error}$, $p_{mis}$, $p_{tos}$, and $\tau_{prob}$ for determining human existence in images were also learned using this training sample set. The estimated values of these parameters are listed in Table VIII.

We designed two schemes to compare our hierarchical handling method with the method which uses the adult image recognition algorithm alone.

—In the first scheme, the subset of the 421 non-adult images, which include 219 non-human images and 202 human images, was taken as the data to test the two adult image recognition methods.
Table IX. The Recognition Results for the First Test Scheme

<table>
<thead>
<tr>
<th>Test images</th>
<th>Recognition methods</th>
<th>Our hierarchical method</th>
<th>The adult image recognition algorithm alone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False rate</td>
<td>Total time</td>
<td>False rate</td>
</tr>
<tr>
<td>Images without human existence</td>
<td>0.46%</td>
<td>4.05s</td>
<td>3.20%</td>
</tr>
<tr>
<td>Images with human existence</td>
<td>9.29%</td>
<td>3.80s</td>
<td>17.33%</td>
</tr>
<tr>
<td>All images</td>
<td>4.99%</td>
<td>7.85s</td>
<td>9.98%</td>
</tr>
</tbody>
</table>

Table X. The Recognition Results for the Second Test Scheme

<table>
<thead>
<tr>
<th>False recognition rate</th>
<th>Human existence</th>
<th>Adult image recognition</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False rate</td>
<td>Total time</td>
<td>False rate</td>
</tr>
<tr>
<td>1.70%</td>
<td>3.65%</td>
<td>5.35%</td>
<td>4.93s</td>
</tr>
</tbody>
</table>

Table IX compares the false recognition rates and the total runtimes of the two methods for the images without human existence, the images with human existence, and all the images in the test subset, respectively. From the table, it is shown that the total runtime of our hierarchical method is much less than that of the method which uses the adult-image recognition algorithm alone. The false recognition rate of our hierarchical method is much less than that of the method which uses the adult-image recognition algorithm alone. The reason why our method obtains more accurate results is that image information, text information, and title information are combined in our method to determine human existence in images.

In the second scheme, the subset of the 411 adult images was taken as the data to test the two methods. The results are shown in Table X. It is shown that while the runtime of our hierarchical method is slightly higher than the runtime of the method that uses the adult-image recognition algorithm alone, the false recognition rate of our method is the same as that of the method which uses the adult-image recognition algorithm alone. The reason for this is that the images that are falsely classified as non-human existence by our human existence detector have very inaccurate skin detection results, and these images are also mistakenly classified by the adult image recognition algorithm.

From the experimental results for the preceding two schemes, it is concluded that our hierarchical method not only decreases the false recognition rate, but also greatly decreases the total runtime for classifying the non-adult images.

A.5 Summary

In our hierarchical method for recognizing Web adult images, images, texts associated with images and titles of Web pages have been fused to determine human existence with a very low computational cost, resulting in an improvement of the performance over using the adult-image recognition algorithm alone.

ACKNOWLEDGMENT

The authors thank Professor Stephen Maybank of Birkbeck College, UK, for his valuable suggestions on the work on adult video recognition.

REFERENCES


**REFERENCES**


