EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Content-based obscene video recognition by combining 3D spatiotemporal and motion-based features

Alierza Behrad[1*], Mehdi Salehpour[1], Meraj Ghaderian[1], Mahmoud Saiedi[2] and Mahdi Nasrollah Barati[1]

Abstract

In this article, a new method for the recognition of obscene video contents is presented. In the proposed algorithm, different episodes of a video file starting by key frames are classified independently by using the proposed features. We present three novel sets of features for the classification of video episodes, including (1) features based on the information of single video frames, (2) features based on 3D spatiotemporal volume (STV), and (3) features based on motion and periodicity characteristics. Furthermore, we propose the connected components' relation tree to find the spatiotemporal relationship between the connected components in consecutive frames for suitable features extraction. To divide an input video into video episodes, a new key frame extraction algorithm is utilized, which combines color histogram of the frames with the entropy of motion vectors. We compare the results of the proposed algorithm with those of other methods. The results reveal that the proposed algorithm increases the recognition rate by more than 9.34% in comparison with existing methods.

Keywords: Obscene video recognition, Content-based video retrieval, 3D spatiotemporal features, Key frame extraction

1. Introduction

Today, the Internet is growing exponentially in different directions, including users, bandwidth, applications, and websites. Nowadays, the Internet has become an essential part of our life, and children are not excluded. Internet provides children many opportunities for learning, research access, socialization, entertainment, and an enhanced communication tool with families while exposing children to potentially negative contents. Because of the fast growth rate of the Internet facilities, the harmful contents on the Internet are growing faster too. Therefore, uncontrolled access to the Internet gives rise to serious social problems.

Content filtering is a commonly used technique by organizations such as schools to prevent computer users from viewing inappropriate web sites or contents. In content filtering techniques, a content is blocked or allowed based on the analysis of its contents not its source. Web contents may include text, image, or video contents. By utilizing content-based filtering, it is possible to block some parts of contents, rather than blocking all web pages or the entire web site.

Video contents have more damaging effect on children and teenagers, among all harmful web contents. Today harmful video contents are employed in different web applications like video files transferring, video chats, live sex, and online videos. Therefore, the recognition of obscene video contents plays an important role in the harmful web contents filtering.

Different methods have been proposed for the task of content-based web filtering; however, most of them have been focused on image or text contents. Recently, a few methods have been proposed for content-based video filtering; however, they mostly employ spatial features like image-based methods. Image-based methods use only the spatial information of single frames for video content analysis and are generally fast. However, video-based approach combines spatial, temporal, and motion-based features for efficient video content analysis and recognition. They are generally more accurate but at the expense of more computational burden.

* Correspondence: behrad@shahed.ac.ir
[1]Faculty of Engineering, Shahed University, Tehran, Iran
Full list of author information is available at the end of the article

In this article, we propose a new approach for the content-based video filtering, which combines different properties of video contents including spatial, spatiotemporal, and motion-based features for robust recognition.

The remainder of this article is organized as follows. In Section 2, existing methods on obscene video recognition are discussed. The proposed features and algorithm for obscene video identification are described in Section 3. Section 4 presents our experimental results, including data collection, training, and test processes. Finally, we conclude the article in Section 5.

2. Methods

Although most of the existing methods have focused on obscene content detection in images [1-3] and texts [4], some efforts have been made for obscene video detection and categorization. Existing method for obscene video detection may be roughly divided into three groups including (1) methods based on spatial information of video frames [5-7], (2) methods based on motion vectors [8,9], and (3) methods based on spatiotemporal features [10-12]. Wang et al. [5] used a three-step method for identifying illicit videos. In the first step, they extracted key frames based on tensors and motion vectors. Then, a cube-based color model was employed for the skin detection. Finally, objectionable videos were recognized by the video estimation algorithm. The method employed only the spatial information of key frames for illicit video recognition. Choi et al. [6] proposed X Multimedia Analysis System (XMAS) for the recognition of obscene video frames. XMAS presented a method for the recognition of obscene videos based on multiple models and multi-class SVM. The system sampled video frames with the rate of 1 frame/s and used MPEG-7 visual descriptors for the feature extraction from images. The method uses only spatial features and its functionality is restricted to MPEG-7 files.

Kim et al. [7] first extracted the frames of a video file and detected shot boundaries or key frames. Then they calculated motion vectors and checked if the frame had a global motion or not. In the case of local motion, the algorithm detected skin segments, and utilized edge moments to classify each frame as an objectionable or a benign frame. The method suffers from using the spatial information of only key frames. It needs also a database for moment matching.

Rea et al. [8] proposed a multimodal approach for illicit content detection in videos. The approach employed visual motion information and the periodicity in the audio stream for illicit video detection. The method assumed that the scene involved only two distinct types of motions: a local homogeneous foreground motion and a global homogeneous background motion. Obviously, real-world motions like zoom/close-up will result in ambiguity.

In [9], a method was presented for detecting the human's reciprocating motion in pornographic videos. The approach extracted motion vectors from the MPEG video stream. The motion vectors were smoothed by vector median and mean filters to remove outliers and small motion vectors. Objectionable videos were then extracted by motion-based features. The method used only motion information for classification. Therefore, the algorithm could not recognize objectionable videos with global motions or videos with no considerable motion.

Jansohn et al. [10] utilized the fusion of motion vectors and spatial features for detecting pornographic video contents. Bag-of-Visual-Words based on the histograms of local patches were used as spatial features. The motion analysis was based on MPEG-4 motion vectors extracted by the XViD codec.

Lee et al., [11] used two models of features for objectionable video classification. The first model utilized features based on single-frame information, and the second feature model was based on the group of frames. The features of two models were classified using two support vector machine (SVM) classifiers. Then the final decision function was utilized to combine the results of two models by using the discriminant analysis. They extended their work [12] to a multilevel hierarchical system, which utilized very similar features for detecting objectionable videos. The method included three phases, which were executed sequentially. In the first phase, initial detection was performed based on hash signatures prior to the download or the play of a video. In the second phase, single frame-based features were utilized for the detection followed by a third phase where the detection was completed by features based on the group of frames reflecting the overall characteristics of the video. Both algorithms extracted video frames periodically to avoid the computational overhead for finding the key frames of a video. This method is not proper for the classification of video episodes with different categories in the same video file.

Zhao et al. [13] studied the key techniques of pornographic image/video recognition algorithms, such as skin detection, key frame extraction, and classifier design in the compressed domain. They extracted shot boundaries by applying a threshold on skin area percentage in the frames, and extracted the proposed features. Finally, the classification was performed by a decision tree.

In [14], Bag-of-Visual-Features was used for nudity detection in video files. The features used to build the vocabulary in this method were simply patches (gray-level values) around the interest points. The method first classified the selected frames independently to nude and

non-nude classes. Then voting algorithm was then utilized to detect nudity in the video file. The method employed only spatial features to decide about whole video content. Also, the use of voting algorithm without the extraction of key frames makes the algorithm unsuitable for the classification of small video episodes with different categories. The algorithm of [15] also used spatial features based on Zernike moments to detect nudity in the video file. The approaches used the global motion in the video frames to group frames and reduce the processing time. The method classifies the input video as obscene if it detects more than five successive obscene frames.

In [16], an agent-based system was developed for the detection of videos containing pornographic contents. The algorithm used color moments and HMM classifier to detect pornographic contents. In [17], an adaptive sampling approach, considering the video duration, was proposed with the objective to increase the detection rate and/or reduce the runtime.

In this article, a new method for the recognition of obscene video contents is presented. In the proposed algorithm, different episodes of a video file starting by key frames are classified independently as obscene or normal. The method employs different shape-based features to differentiate between skin regions of obscene and non-obscene videos. We utilize different novel features for obscene video content recognition, including spatial, spatiotemporal, and motion-based features. To extract spatiotemporal features, we employ a novel method based on the 3D skin volume and new concept of the relation tree to find the spatiotemporal relationship between the skin regions in consecutive frames. Also to increase the efficiency of the proposed motion-based features, we propose a new method for key frame extraction that combines color histogram of the frames with the entropy of motion vectors.

3. Proposed algorithm

Figure 1 shows the block scheme of the proposed algorithm. As it is shown in the figure, the algorithm has three stages, including (1) preprocessing, (2) feature extraction, and (3) classification. The algorithm starts by the detection of key frames. When a key frame is detected, the information of video frames is extracted for about 4 s after the key frame and skin regions in video frames are extracted. At the second stage of the algorithm, the proposed features are extracted from binary skin images. Three sets of features are proposed for the classification of video episodes as follows.

- Features based on the information of single frames.
- Features based on 3D STV.
- Features based on motion and periodicity characteristics.

Features based on the information of single frames extract features from individual frames of the video. The method is fast for feature extraction; however, it uses only the spatial information of single frames for video content analysis. Features based on 3D STV consider not only the spatial characteristics of the individual frames, but also their temporal variation over video frames. To extract spatiotemporal features in a video

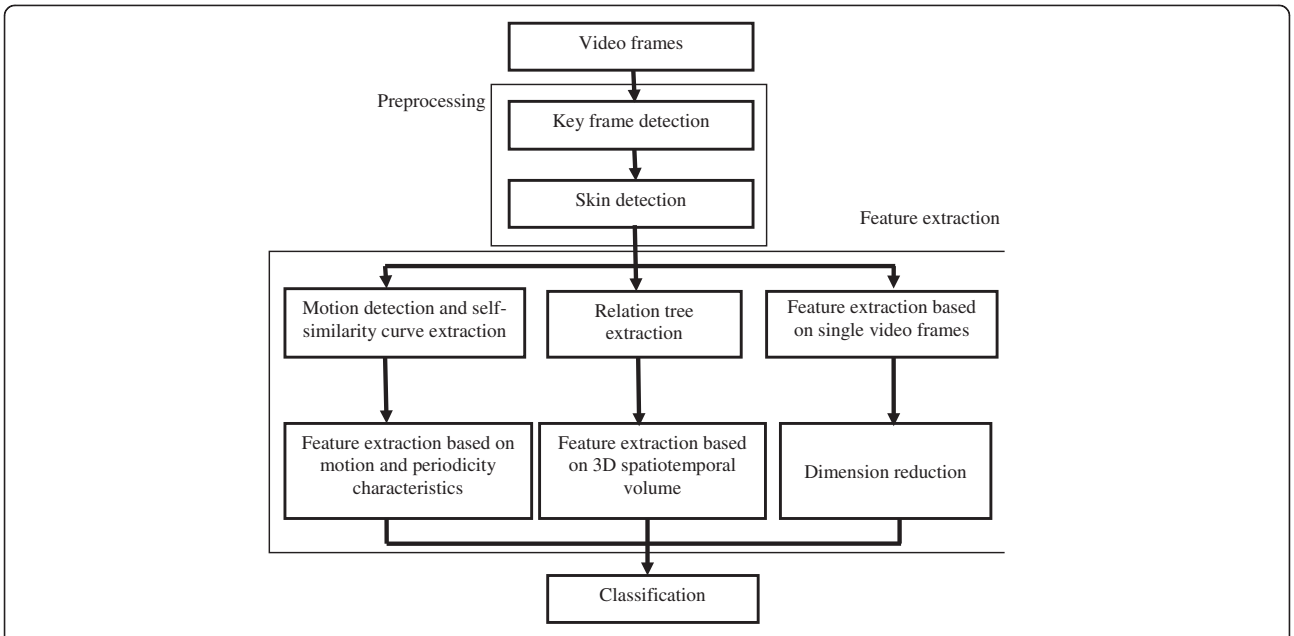

Figure 1 Block scheme of the proposed algorithm.

episode, we construct connected components' relation tree, which shows the spatiotemporal relationship between the skin regions in consecutive frames. Motion is a key feature representing temporal characteristics of videos. Periodicity of motion is the main characteristic of obscene videos, which can be used as another feature for the classification of obscene and normal videos. However, when there is no motion in the scene or when the scene includes a global motion, motion-based features are not reliable for periodicity measurement. Therefore, we consider the validity of motion-based features for more efficient classification.

At the last step of the algorithm, all the features are combined and the video episode is classified using an SVM classifier.

3.1. Preprocessing

The main goal of preprocessing step in the proposed algorithm is to divide the video file into video episodes by the detection of key frames. Each video episode can be classified independently as obscene or non-obscene. In addition, skin regions are extracted in the preprocessing stage. Since the skin detection algorithm may not detect skin pixels completely, we apply necessary post-processing techniques for noise handling.

3.1.1. Key frame detection

Since various video parts may contain different contents, the proposed algorithm is devised to classify different episodes of a video file independently as obscene or non-obscene. For this purpose, we need to divide a video file into video episodes. In addition, due to massive video data, video summarization is a necessary stage to organize video data and implement a meaningful rapid navigation of video. Video summarization is the process of creating a new representation of video data that is much shorter than original video data and information is preserved as much as possible. Video summarization algorithms generally aim at finding events with more valuable information in the video streams, reducing the network load and preparing useful data for the classification.

Key frame detection is the mostly used technique for video summarization. By the extraction of key frames, first, a video file is divided into a collection of video episodes that can be examined separately. Second, since we use only the information of video frames for the time interval of 4 s after key frames, the computation burden of the algorithm is reduced. Different methods have been proposed for key frame extraction, including color-based methods [18], methods based on motion vectors [19,20], object-based techniques [21,22], and methods based on feature vector space [23,24] to name a few.

Our method for key frame extraction combines color histogram of the frames with the entropy of motion vectors. The algorithm includes two successive steps. In the first step, color histogram of frames is employed as follows.

- Color histograms of video frames are calculated.
- Normalized cross correlation coefficients between histograms of consecutive frames are calculated.
- Local minimums of cross correlation coefficients are identified.
- Key frames are detected by applying an appropriate threshold to cross correlation coefficients.

In the case of videos with poor illumination, color histogram may generate myriad of key frames without any changes in the scene or motion information. In addition, we use motion features for the classification of video episodes, which means key frames should reveal a change in motion information as well. Therefore, motion information in the second step of key frame detection algorithm is employed. The purpose of this step is to eliminate some key frames that reveal no change in motion information. We use the entropy of motion vectors to extract motion information in two consecutive frames. Motion vectors are calculated using block matching algorithms for all blocks of video frames. Two-dimensional motion vectors are then mapped to an intensity image where the intensity values are calculated using the following equation

$$I(x,y) = (2R+1)(d_x + R) + d_y + R \tag{1}$$

where (d_x, d_y) is the vector representing the motion of the pixel (x, y). It is assumed that square areas with the size of $(2R + 1) \times (2R + 1)$ are used as search regions in the block matching approach.

To extract motion information, co-occurrence matrix for image I is calculated. Assuming that the input frames contain two different areas, including background (non-skin) and foreground (skin) areas and their motion vectors are separated by threshold t, the co-occurrence matrix is divided into four quadrants, which repre sent background-to-background (BB), background-to-foreground (BF), foreground-to-background (FB), and foreground-to-foreground (FF) regions. The entropies of the quadrants are calculated using the following equations [25]:

$$H_{BB}(t) = -\sum_{i=0}^{t}\sum_{j=0}^{t} p_{BB}(i,j)\log\, p_{BB}(i,j) \tag{2}$$

$$H_{BF}(t) = -\sum_{i=0}^{t}\sum_{j=t+1}^{L-1} p_{BF}(i,j)\log\, p_{BF}(i,j) \tag{3}$$

$$H_{FF}(t) = -\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{FF}(i,j)\log\ p_{FF}(i,j) \tag{4}$$

Then global, local, and joint entropies that show the motion information of a frame are calculated as follows:

$$H_{LE}(t) = H_{BB}(t) + H_{FF}(t) \tag{5}$$

$$H_{LE}(t) = H_{BB}(t) + H_{FF}(t) \tag{6}$$

$$H_{JE}(t) = H_{FB}(t) + H_{BF}(t) \tag{7}$$

$$H_{GE}(t) = H_{FB}(t) + H_{BF}(t) + H_{BB}(t) + H_{FF}(t) \tag{8}$$

$$H_{LEM} = \max_{t=1}^{L-1} H_{LE}(t) \tag{9}$$

$$H_{JEM} = \max_{t=1}^{L-1} H_{JE}(t) \tag{10}$$

$$H_{GEM} = \max_{t=1}^{L-1} H_{GE}(t) \tag{11}$$

where H_{GEM}, H_{LEM}, and H_{JEM} are maximum global, local, and joint entropies, respectively. A key frame should reveal a considerable change in motion information. Therefore, we define motion information difference (MID) between two consecutive frames i and $i-1$ as:

$$MID = \left|H^{i}_{GEM} - H^{i-1}_{GEM}\right| + \left|H^{i}_{JEM} - H^{i-1}_{JEM}\right| + \left|H^{i}_{LEM} - H^{i-1}_{LEM}\right| \tag{12}$$

where H^{i}_{LEM}, H^{i}_{JEM}, and H^{i}_{GEM} are maximum local, joint and global entropies for frame i, respectively, and H^{i-1}_{LEM}, H^{i-1}_{JEM} and H^{i-1}_{GEM} are maximum local, joint and global entropies for frame $i-1$, respectively. By employing MID values, the key frames extracted by the first step of the algorithm are further refined to extract more reliable key frames.

3.1.2. Skin detection

Majority of obscene videos contain large volume of skin region. Therefore, skin regions are an obvious cue for the recognition of obscene videos. Several methods have been proposed to detect skin pixels in image [26-29]. In pixel-based approaches, each pixel is classified as skin or non-skin pixels individually and independently from its neighbors [26,27]. In contrast, region-based approaches take spatial arrangement of pixels into account during the detection stage [28,29]. Much of the existing work on skin detection has used a mixture of Gaussian models for skin extraction. A mixture of Gaussian models is expressed as the sum of Gaussian kernels as follows

$$P(\mathbf{x}) = \sum_{i=1}^{N} \omega_i \frac{1}{(2\pi)^{\frac{3}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} \tag{13}$$

where $\mathbf{x}$ is the color vector, Σ_i are diagonal covariance matrices, and μ_i are the mean vectors. The contribution of the ith Gaussian function is specified by ω_i. In [30], several algorithms for skin detection in objectionable videos were compared. It was shown that the mixture of Gaussian models is a proper choice for skin detection in objectionable videos. The implementation of the mixture of Gaussian models using a lookup table makes the skin detection algorithm proper for real-time applications as well. We use the method presented in [26] which employs two separate mixture models for the skin and non-skin classes. The method exploits 16 Gaussians in each model and extracts skin pixels by applying threshold on the skin likelihood which is defined as follows

$$L(\mathbf{x}) = \frac{P_{skin}(\mathbf{x})}{P_{non-skin}(\mathbf{x})} \tag{14}$$

where $L(\mathbf{x})$ is the skin likelihood. To remove erroneous skin pixels and to have uniform skin region, the following post-processing stage is applied to the resultant binary skin image.

- Morphological opening operator is applied to remove small connected components (skin regions) in the image.
- Pixels with less than four skin pixels in their 3×3 neighborhood are removed.
- Morphological closing operator is applied to merge nearby skin regions.
- Holes in skin regions are filled.

Figure 2 shows the results of different stages for the skin detection algorithm.

3.2. Feature extraction

Feature extraction has a great impact on the performance of the video recognition system. We use three different sets of features for the recognition of obscene videos, namely features based on the information of single frames, features based on 3D STV, and features based on motion and periodicity characteristics.

These features are extracted for each episode of video starting by a key frame. For this purpose, after extracting key frames, frames of a video episode with the duration of about 4 s are extracted. Then after applying skin detection algorithm, the required features are calculated.

To extract volume-based features, connected components (skin regions) of the skin image are extracted and their spatiotemporal relationship and arrangement in the consecutive frames are evaluated. For this purpose, we propose connected components' relation tree in successive frames that are explained in the next section.

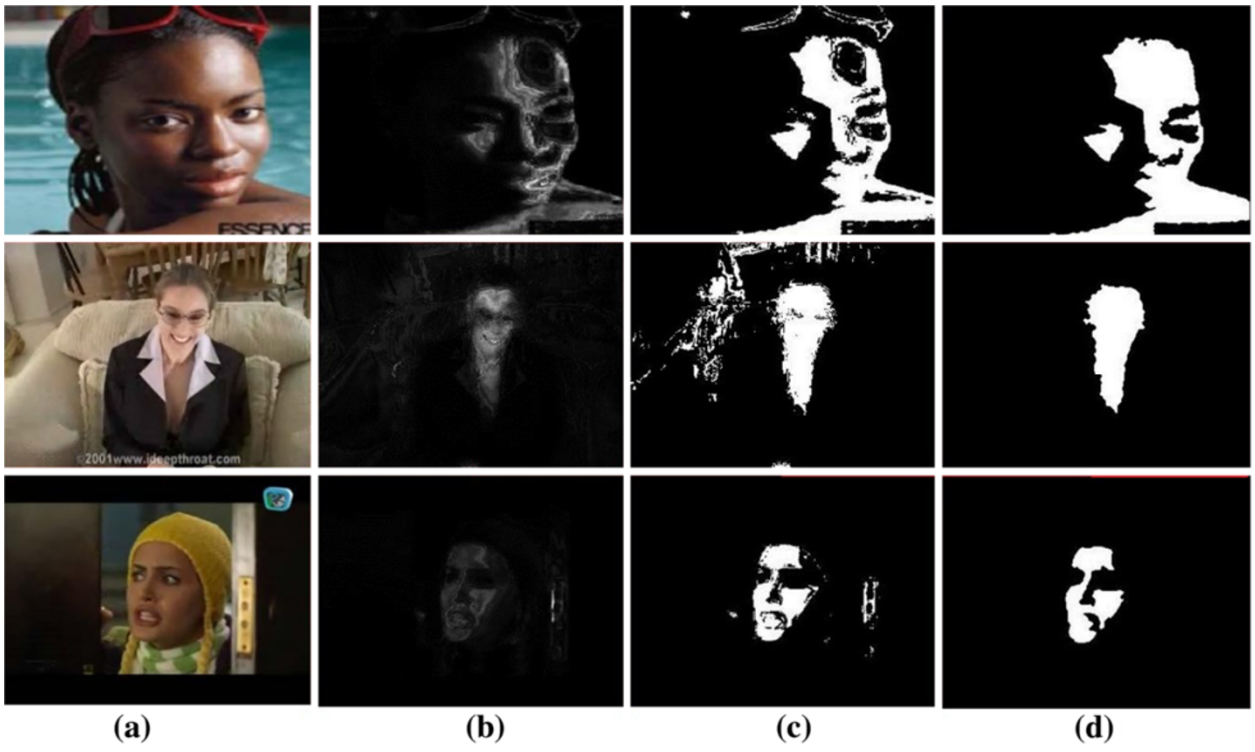

Figure 2 Results of skin detection algorithm. (**a**) Original images, (**b**) likelihood images, (**c**) binary skin images, (**d**) skin images after applying post-processing stage.

3.2.1. Connected components' relation tree

We use the relation tree to find the spatiotemporal relationship between the skin regions in consecutive frames. The relation tree is used to extract the features based on 3D STV. For this purpose, first the skin regions of consecutive frames are labeled and three largest regions are selected to reduce the computational burden. To enhance the robustness of the algorithm, small connected components are eliminated. Consequently, some frames may have less than three connected components.

Algorithm for the construction of the relation tree starts by finding the first frame which must contain at least one connected component. The relationship between connected components is then calculated in subsequent frames and the relation tree is constructed.

Figure 3 shows an example of a relation tree for four successive frames. Each node in this directional tree is shown by a circle, representing a connected component or a skin region. Directional link between two nodes represents a relationship or an overlap between two nodes and the cost of the link represents the amount of overlap between two nodes (connected components) in terms of pixel. Three kinds of nodes are defined in the relation tree as follows

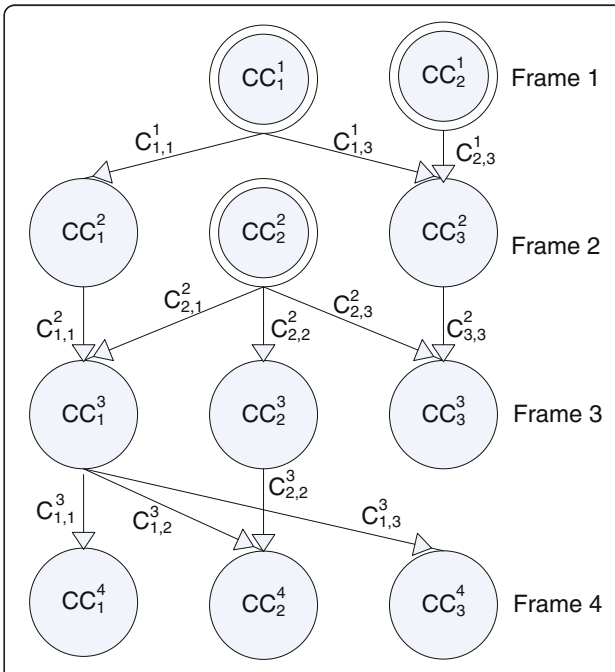

Figure 3 Four steps of the relation tree's progress.

- *Parent node*: a node that does not have any predecessor. Nodes CC_1^1, CC_2^2, CC_2^1 are parent nodes in Figure 3.
- *End node*: a node that does not have any successor. Nodes CC_3^3, CC_1^4, CC_2^4, CC_3^4 are end nodes in Figure 3.
- *Intermediate node*: nodes that relate parent nodes to end nodes.

To construct the relation tree between two consecutive frames, skin regions or connected components in the current frame are compared with the connected components in the next frame. If two skin regions CC_i^j and CC_k^{j+1} in two subsequent frames has overlap, then the link $l_{i,k}^j$ with the cost of $C_{i,k}^j$ is added to the tree, where $C_{i,k}^j$ is the number of overlapped pixels between two skin regions. Pseudocode for the construction of the relation tree between two consecutive frames is shown in Figure 4.

A path is defined as a sequence of nodes CC_1, CC_2, ..., CC_k and their related links, where CC_1 is a parent node, CC_k is an end node, and each intermediate node CC_i is the successor of CC_{i-1}. Cost of a path is defined as the sum of costs for all the links in the path.

Notations:
N_k : Number of connected components in frame k
CC_i^k : i^{th} connected component in frame k
$l_{i,j}^k$: Link between CC_i^k and CC_j^{k+1}
$C_{i,j}^k$: Cost of link $l_{i,j}^k$
P^k : All paths formed until frame k
P_i^k : Paths that ends to CC_i^k

Algorithm:

```
For i:=1 to N_k
    first_link:=1
    For j:= 1 to N_{k+1}
        If ( CC_i^k ∩ CC_j^{k+1} ≠ ∅ )
            Add link l_{i,j}^k with C_{i,j}^k = CC_i^k ∩ CC_j^{k+1} to tree
            If( first_link==1)
                Add link l_{i,j}^k to P_i^k to form P_j^{k+1}
                first_link:=0
            Else
                Copy P_i^k to temporary paths Pt_j^k
                Add link l_{i,j}^k to Pt_j^k to form new paths Pt_j^{k+1}
                Add Pt_j^{k+1} to P^{k+1}
            Endif
        Endif
    Endfor
Endfor
```

Figure 4 Pseudocode for creating relation tree between two consecutive frames.

After creating the relation tree, the optimal path, which is defined as a path with the maximum number of nodes, is selected. If a few paths have the maximum number of nodes simultaneously, the path with maximum cost is selected as the optimal path. The optimal path is used for the construction of 3D STV and feature extraction.

3.2.2. Features based on the information of single frames

Although skin regions are one of the important characteristics of obscene images and videos, some normal video frames may also have a significant percentage of skin regions such as face regions. Therefore, suitable features should be extracted from skin regions. For this purpose, we use features based on the shape of skin regions for the classification. The first group of the proposed features is based on the information of single frames. These features that are extracted for all frames in the video episode include

- the area of the largest skin region in the frame;
- hydraulic factor which is defined as the area to perimeter ratio of the largest skin region;
- solidity which is defined as the area of the largest skin region to the area of its bounding convex hull;
- compactness factor which is defined as the area of the largest skin region to the area of bounding box for all skin regions in the frame;
- minor to major axis ratio of the ellipse that has the same normalized second central moments as the largest skin region in the frame;
- equivalent diameter of the circle with the same area of skin regions in the frame.

Since these features are calculated for all existing frames in the video episode, the size of features is large. Hence, we utilize the principal component analysis (PCA) approach to reduce the features' dimension [31]. In the PCA approach, mean vector and covariance matrix are calculated for all existing data in the database.

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{15}$$

$$\hat{X}_i = X_i - \overline{X} \tag{16}$$

$$W = \left[\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N\right] \tag{17}$$

$$C = \frac{1}{N}\sum_{i=1}^{N} \hat{X}_i \hat{X}_i T = \frac{1}{N} W W^T \tag{18}$$

where $\overline{X}$ and C are mean vector and covariance matrix. Then PCA is applied to the covariance matrix C, and M largest principal components are used for the feature extraction as follows

$$Y_i = \left(X_i - \overline{X}\right)^T D \tag{19}$$

where Y_i are the calculated features, and D is the matrix of M principle vectors. We experimentally use the value of 20 for M.

3.2.3. Features based on 3D STV

The frame-based features, which are extracted independently for each frame, are spatial features that do not show temporal characteristics of the skin regions. STVs unify the analysis of spatial and temporal information by constructing a volume of data in which consecutive frames are stacked to form a third, temporal dimension. After the extraction of the optimal path, the connected components of the optimal path are extracted. Then the extracted connected components are stacked over each other to construct a 3D STV. The volume shows the spatial characteristics of connected components in the optimal path and their temporal variation. Two groups of shape-based features are extracted from the volume. The first group includes six features as follows.

- The volume of the STV which is defined as the number of skin pixels in all connected components in the volume.
- Volume solidity (VS) which is defined as the ratio of pixels in convex hull volume to the number of skin pixels in STV. To obtain convex hull volume, we extract bounding convex hull for all connected components in the path, and VS is calculated using the following equation:

$$VS = \frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} S_i} \tag{20}$$

where A_i and S_i are the areas of ith connected component in the optimal path and its convex hull, respectively, and N is the number of connected components in the optimal path.

- Volume hydraulic factor (VHF) that is defined as the volume to surface ratio of STV as follows

$$VHF = \frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} P_i} \tag{21}$$

where A_i and P_i are the area and perimeter of ith connected component in the STV, respectively, and N is the number of connected components in the optimal path.

- Equivalent sphere diameter which is defined as the diameter of a sphere with the same volume as the STV volume.
- Volume compactness which is defined as the ratio of STV volume to the volume of rectangular parallelepiped bounding the STV.
- Average diameter of circles with the same areas of connected components in the optimal path.

To extract second group of features, we first map all the connected components in the STV to a single image called optimal path map image (OPMI). OPMI is calculated using the following equation:

$$OPMI(i,j) = \begin{cases} 1 \, if \sum_{k=1}^{N} STV(i,j,k) \neq 0 \\ 0 \, o.w. \end{cases} \tag{22}$$

where N is the number of connected components in the STV, and STV(i,j,k) is the value of STV with the spatial coordinate of (i, j) and the temporal coordinate of k. STV(i,j,k) is '1', if the pixel with the coordinate of (i,j,k) is a skin pixel, otherwise its value is set to '0'. After calculating OPMI, the connected component in OPMI is extracted and the second group of volume features is calculated as follows

- OPMI solidity which is defined as the ratio of the connected component area in OPMI to the area of its bounding convex hull.
- OPMI hydraulic factor which is defined as the area to perimeter ratio of the connected component in OPMI.
- OPMI compactness factor which is defined as the ratio of the connected component area to the area of its bounding box.
- Minor to major axis ratio of the ellipse that has the same normalized second central moments as the connected component in the OPMI.
- Diameter of the circle with the same area of the connected components in the OPMI.

In obscene videos, there is a considerable volume of skin pixels in consecutive frames and generally with periodic motion. Therefore, the connected component in OPMI image is larger and generally not very scattered. However, in normal videos, the connected component is smaller or scattered. The OPMI features enhance discrimination property of the proposed features.

3.2.4. Features based on motion and periodicity characteristics

Motion is a key feature representing temporal characteristics of videos. Motion features have been used in

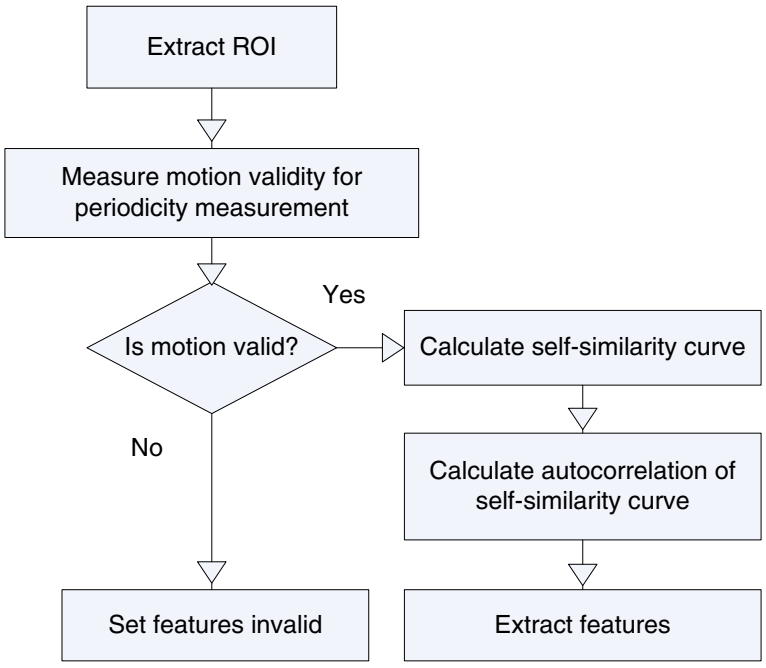

Figure 5 Feature extraction using motion and periodicity characteristics.

different applications like video retrieval [32], action recognition [33], and human identification [34] to name a few. Periodicity of motion and its rate are the key property of obscene videos, which can be used as a cue for the feature extraction and classification.

Recently, some algorithms have been proposed to detect periodic motion and its features to overcome the problems of traditional human motion analysis approaches [35-39]. In [35], periodic motion was defined as repeating curvature values along the path of motion. The method detected periodic motion using spatiotemporal (ST) surfaces and ST-curves. The projected motion of an object generates ST-surface. ST-curves were detected on the ST-surfaces, providing an accurate description of the ST-surfaces. Curvature scale-space presentation of the ST-curves was then used to detect intervals of repeating curvature values. Briassouli and Ahuja [36] provided a method based on time-frequency analysis of the video sequence. Cheng et al. [37] introduced a feature descriptor to classify different kinds of sports with periodic motion. The method utilized motion vectors in the horizontal and vertical directions as the basis to extract periodicity features.

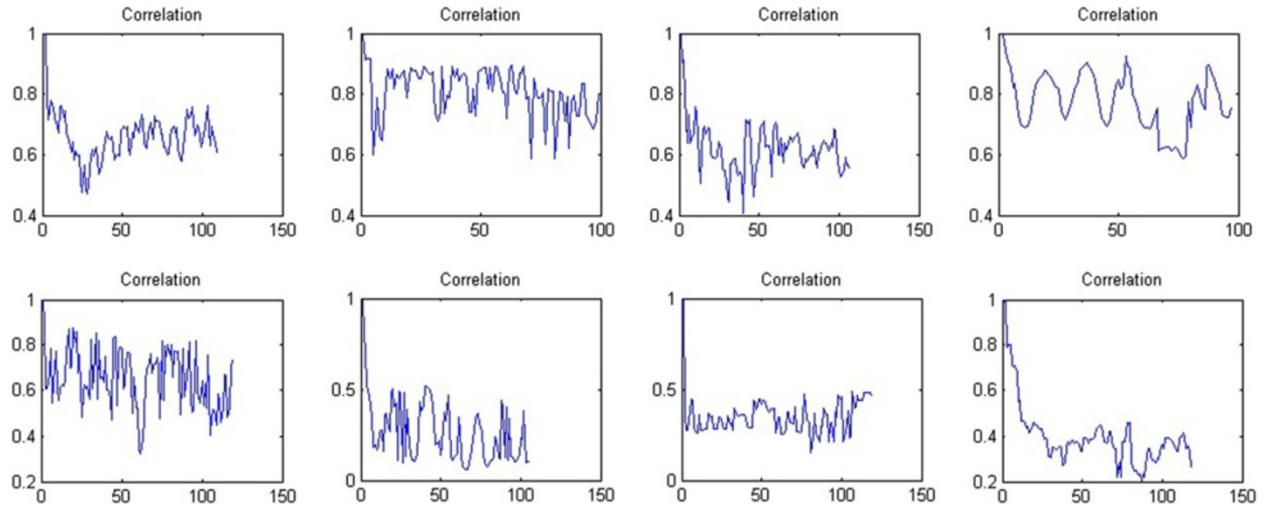

Figure 6 Self-similarity curves for some obscene video episodes.

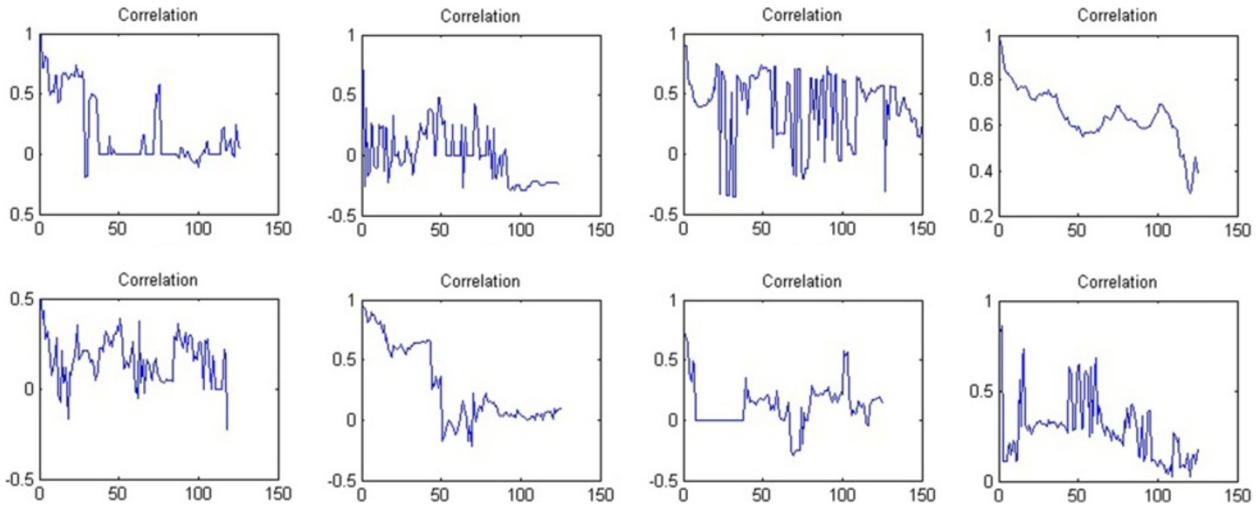

Figure 7 Self-similarity curves for some non-obscene video episodes.

Cutler and Davis [38] dealt with the recognition and analysis of periodic motions. In their method, first moving objects were segmented. Then for the recognition of objects' period, the segmented objects in each frame were aligned using object centers and all objects were resized to have the same sizes. The similarity measure and autocorrelation of the objects were then used to estimate the periodicity of the motion. Tong et al. [39] extracted local motion in the consecutive frames and determined the object area using the motion segmentation algorithm. The method calculated the mean squared of motion vectors in each frame and obtained the motion curve. The local maximums of the autocorrelation of the motion curve were then used to extract periodicity features by fitting proper Gaussian functions.

Some of the mentioned methods utilize motion vectors or features from motion vectors for the recognition of periodic motion. The main problem of these methods is their low accuracy in the calculation of motion vectors because of non-rigid and flexible motion of the human body. In addition, the computation burden of these algorithms is high. Another group of algorithms is based on the self-similarity measure of moving objects in the consecutive frames where the correlation of intensity values is the mostly used method for the self-similarity measure. The autocorrelation of intensity values is insensitive to motion outliers and less affected by illumination change.

We use a method based on the similarity measure to extract features representing the periodicity of motion and its specification. However, most of the mentioned

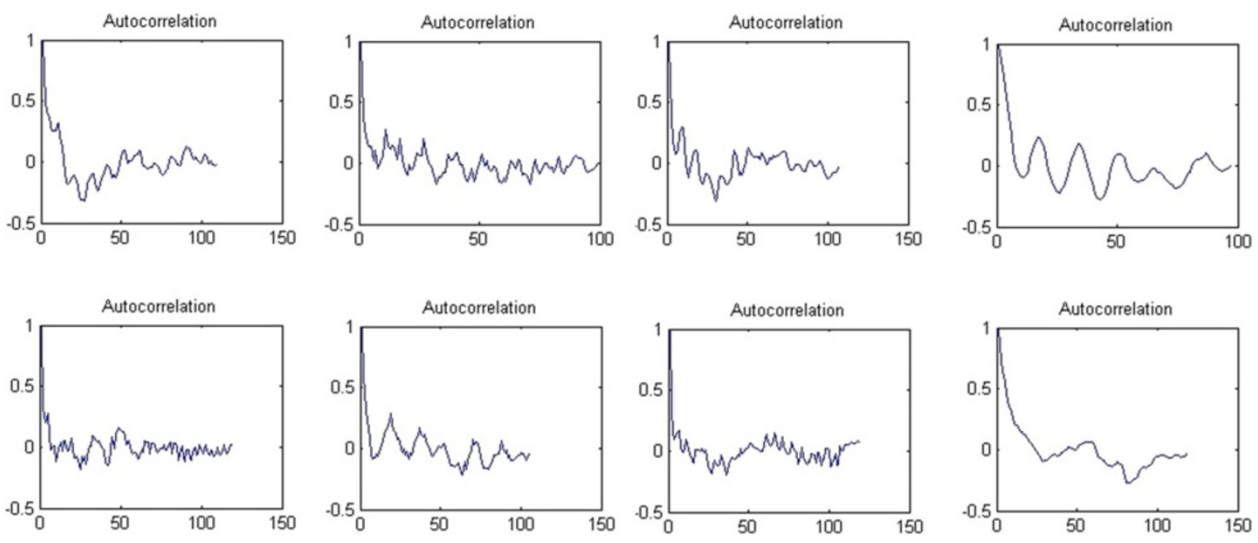

Figure 8 The autocorrelation of curves in Figure 6.

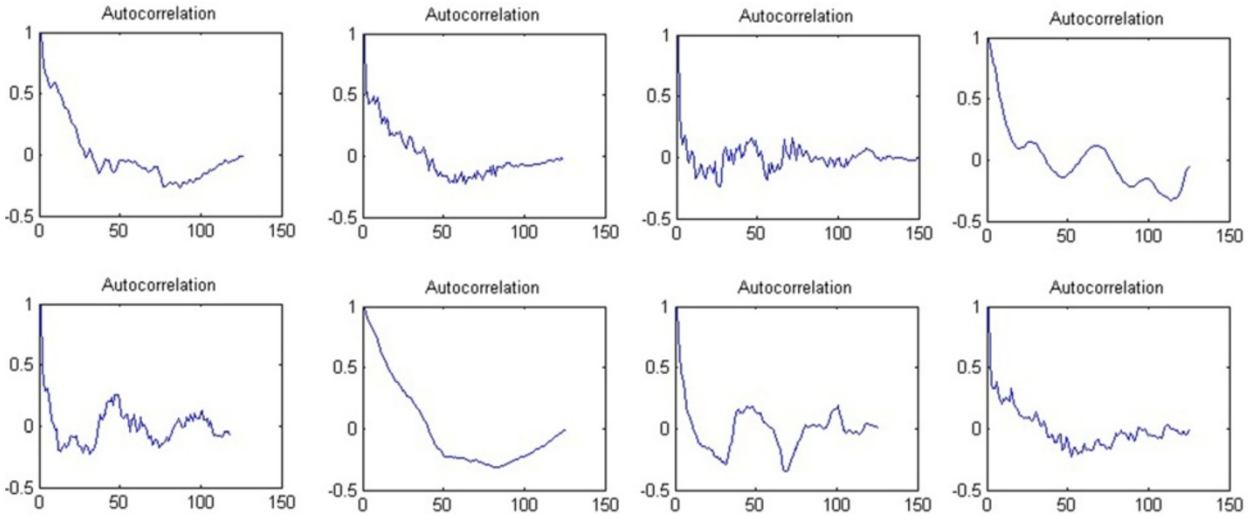

Figure 9 The autocorrelation of curves in Figure 7.

methods measure the periodicity of motion in the restricted situations like stationary camera and known environments, which are not applicable to our real-world application. To handle this problem, we use the algorithm depicted in Figure 5, the description of its different stages are described as follows.

3.2.5. Extracting ROI

The first stage in the periodicity analysis is the extraction of region of interest (ROI). In most of the algorithms, moving objects in the scene are used for the analysis; however, the method fails when the camera is also moving. We use skin region as ROI for periodicity analysis in our algorithm. We first extract skin regions in the consecutive frames using the method described in Section 3.1.2. Then by applying proper morphology operators, very close connected components (skin regions) are merged and holes are filled. Finally, largest connected component is kept and other connected components are removed.

3.2.6. Motion validity measurement

Motion- and periodicity-based features are meaningful when there is a considerable motion in the video episode starting by a key frame. However, some of both obscene and non-obscene video episodes may be static. In addition, when the camera is also moving, the motion-based features are mostly affected by camera motion. To deal with this problem, we measure validity of motion in consecutive frames and extract motion-based features only when the motion is valid for the periodicity measurement.

To calculate motion validity, we first extract moving pixels in two consecutive frames using the image subtraction algorithm. If moving regions outside the skin regions in a frame is larger than 50% of skin regions, the frame is considered as a frame with camera motion. If the number of video frames with camera motion is less than a predefined threshold, the motion validity flag is set to calculate the periodicity-based features otherwise all the periodicity-based features are set to zero.

3.2.7. Self-similarity curve calculation

We use self-similarity curve to detect periodicity of skin ROI in the proposed method. The self-similarity curve is defined as

$$S_{t_1}(k) = Sim(ROI(k), ROI(t_1)) \tag{23}$$

where $S_{t_1}(k)$ is the self-similarity curve, k is the temporal index, ROI(k) is the image of ROI in frame k, and Sim function is an image similarity metrics. When the motion of ROI image is periodic, the self-similarity curve is also periodic with the same period. Different image similarity metrics may be used for the similarity curve extraction.

$$S(k) = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(ROI(i,j,0) - \overline{ROI(0)}\right)\left(ROI(i,j,k) - \overline{ROI(k)}\right)}{\sqrt{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(ROI(i,j,0) - \overline{ROI(0)}\right)2\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(ROI(i,j,k) - \overline{ROI(k)}\right)^2}} \tag{24}$$

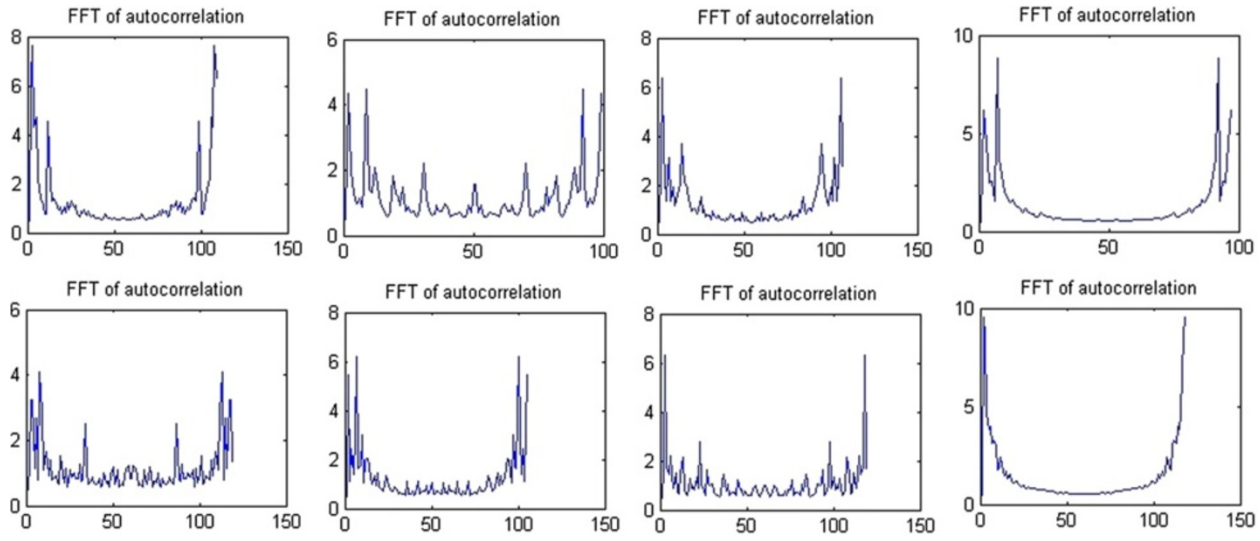

Figure 10 Amplitude spectrum of autocorrelation curves in Figures 8.

We use normalized cross correlation for the self-similarity curve extraction as follows

Since the ROI image is a binary image, the calculation of $S(k)$ is computationally inexpensive. Figures 6 and 7 show the samples of self-similarity curves for some obscene and non-obscene video episodes, respectively.

3.2.8. Reliability of the self-similarity curve

In videos with small skin area, skin regions may not be detected in some frames. In this case, the self-similarity curve is short or oscillatory, which results in non-reliable periodicity features. For this purpose, we define the reliability factor (RF) for the self-similarity curve as:

$$RF = (N_t - K_f)N_s/{N_t}^2 \tag{25}$$

where N_t is the total number of frames in the video episode, N_s is the number frames with skin region, and K_f is the temporal index of first frame with skin region.

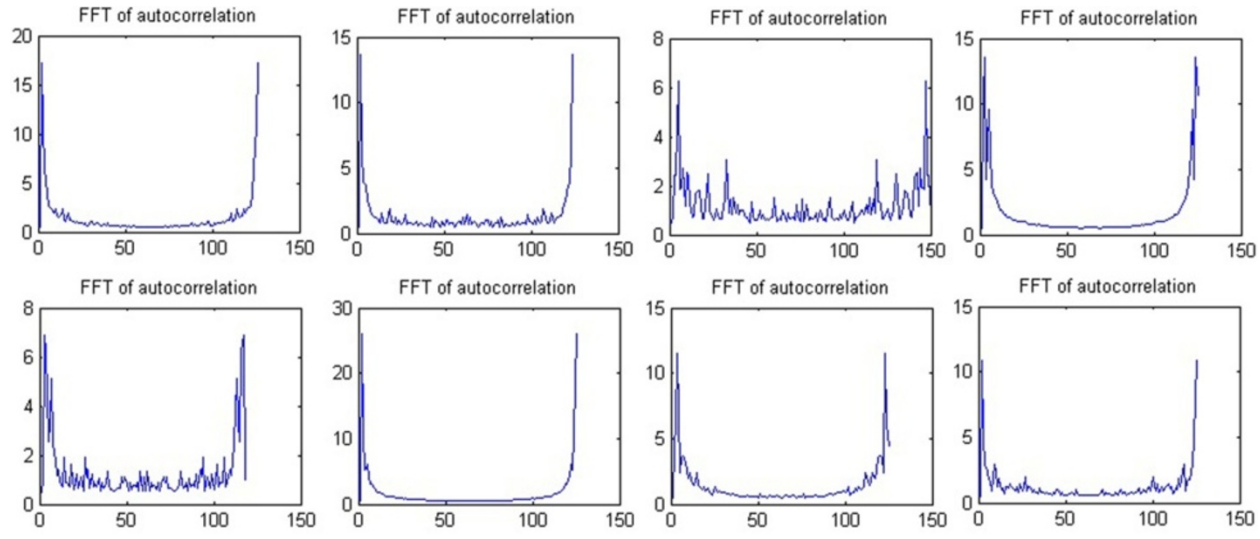

Figure 11 Amplitude spectrum of autocorrelation curves in Figures 9.

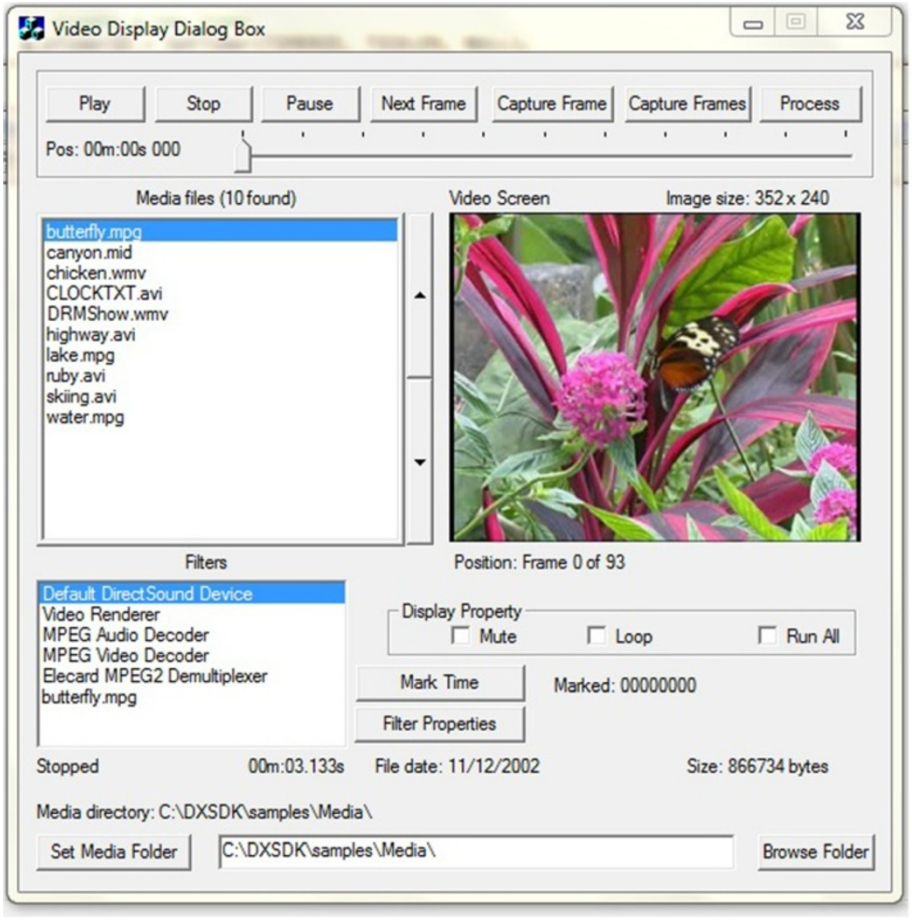

Figure 12 The implemented application program for the test of the proposed algorithm.

Table 1 Specification of the collected dataset

Category			Duration (minute)
1	Obscene	Animation (porn)	329.49
2		Animal sex	150.29
3		Bad illumination	15.87
4		Heterosexual	2200.28
5		Dildo	603.97
6		Gay	434.42
7		Lesbian	59.12
8		Nude	38.25
9		Porn with cloth	16.38
10		Semi-porn	9.01
11	Normal	Animation (non-porn)	338.17
12		Movie	5263.55
13		Music video	35.77
14		Iranian movie	584.94
15		Short-time video clips	425.3
16		Low-resolution video clips	439.11

When the entire frames in the video episode contain skin area, the value of RF is 1.

3.2.9. Autocorrelation of self-similarity curves

As shown in Figures 6 and 7, the self-similarity curves are noisy, and it is difficult to extract proper features directly. To handle this problem, we calculate the autocorrelation of self-similarity curves after subtracting mean value as follows

$$\rho(k) = \sum_{j=0}^{N_t-1} \hat{S}(j)\hat{S}(j+k) \tag{26}$$

where $\hat{S}$ are self-similarity values after removing mean value and N_t is the total number of frames in the video episode. For a periodic signal, the autocorrelation signal is also periodic. Figures 8 and 9 show the autocorrelation of self-similarity curves depicted in Figures 6 and 7, respectively.

Table 2 RR of the proposed algorithm for different experiments

SVM core	RR						
	Exp. #1 (%)	Exp. #2 (%)	Exp. #3 (%)	Exp. #4 (%)	Exp. #5 (%)	Average (%)	SD (%)
RBF	74.50	78.30	82.50	73	76.40	76.94	3.69
Linear	93.20	95.40	98	96.30	94.30	95.44	1.84
Quadratic	83.50	85.67	86.67	87.85	88.60	86.46	1.99
Polynomial	85.50	89.30	87.80	82.60	81.80	85.40	3.23

3.2.10. Feature extraction

To extract motion-based features, we calculate the Fourier transform of autocorrelation coefficients and extract peaks in the amplitude spectrum. Figures 10 and 11 plot amplitude spectrum of autocorrelation curves in Figures 8 and 9, respectively. We use six features for periodicity measurement as follows.

- Frequency of the largest peak in the amplitude spectrum.
- Amplitude of the largest peak in the amplitude spectrum.
- Frequency of second largest peak in the amplitude spectrum.
- Amplitude of second largest peak in the amplitude spectrum.
- Motion validity flag
- RF of the self-similarity curve

3.3. Classification

We use SVM classifier [40] for the classification of video episodes. SVM classifiers are based on the concept of decision planes that define decision boundaries. The original SVM classifier was a linear classifier. However, nonlinear kernel functions were utilized to extend SVM capability for nonlinear classification [41]. The proposed feature vector is a 37-dimensional vector with the following elements.

- Features based on the information of single frames: 20 elements.
- Features based on 3D STV: 11 elements.
- Features based on motion and periodicity characteristics: 6 elements.

We tested SVM classifier with different kernel functions which the results are reported in the following section.

4. Experiments

The proposed algorithm was implemented using a Visual C++ program and tested with different obscene and non-obscene videos. To extract frames' data for different video formats, we implemented an application program which is based on the Microsoft DirectX SDK's. Figure 12 shows a view of the implemented software for testing the proposed algorithm. The implemented software is capable of encoding different video formats and includes suitable interfaces for selecting and testing whole or different parts of the selected video files.

In order to evaluate the proposed algorithm, a large volume of video files were collected by random web surfing. The database includes 1,060 video files with a total duration of 10943.92 min, where 3857.08 min belong to the obscene video category, and 7086.84 min are normal videos. We divided obscene and non-obscene videos into different categories. Table 1 shows different categories and their durations for the collected database.

After applying key frame detection algorithm, we randomly select 2,000 episodes of obscene videos and 2,000 episodes of normal videos for the evaluation of the proposed algorithm. We use 700 normal and 700 obscene video episodes for the training and the remaining 2,600 episodes for the test.

Table 2 shows the recognition rate (RR) of the proposed algorithm using different SVM cores for five different experiments. We use different video episodes for the test and train of each experiment. Table 2 shows the

Table 3 Execution time and the processing frame rate of the proposed algorithm

Frame size	Number of frames	Execution time (s)	Processing frame rate (frames/s)
240*352	1768	50.4	35.08
128*128	1240	28	44.29
240*320	179	4.68	38.25
288*352	33643	799.3	42.09
352*640	4500	186.4	24.14

Table 4 Specification of the processing unit

Component	Specification	Component	Specification
CPU	Intel Core Due 2	RAM	2 GB
CPU frequency	2.4 GHz	OS	Windows XP
Hard capacity	150 GB	Compiler	Visual C++ 6.0

average RR of 95.44% for the proposed algorithm with linear SVM kernel. The results of Table 2 show that the proposed features are linearly separable for obscene and normal videos.

The proposed algorithm is fast and can process video files in real time. Table 3 illustrates the execution time and the processing frame rate of the algorithm for various video files with different frame sizes. The execution time in Table 3 includes all stages of the proposed algorithm. To measure the execution time, we used a laptop and its specifications are shown in Table 4.

To show the effect of individual features on the accuracy of the proposed algorithm, we tested the proposed algorithm by removing some features elements. Table 5 shows the recognition of the proposed algorithm when various features elements are removed. The results of Table 5 show that features based on the information of single frames and 3D STV have the major effects on the accuracy of the proposed algorithm. When the camera is moving, the periodicity-based features are not useful. Furthermore, some of the obscene videos may not have periodic motions; therefore, features based on the motion and periodicity characteristics have less impact on the accuracy of the proposed algorithm.

Table 6 Average RR for the implemented methods

Algorithm	A (%)	B (%)	C (%)	D (%)	E (%)	F (%)	G (%)
RR	80	79.2	86.1	82.9	64.73	81.4	76.7

To compare the results of the proposed algorithm with those of other methods, we implemented the following algorithms:

Algorithm A: Hierarchical system for objectionable video detection [12].
Algorithm B: High performance objectionable video classification system [11].
Algorithm C: Adult image filtering for web safety with SVM classifier [42].
Algorithm D: Adult image filtering for web safety with KNN classifier [42].
Algorithm E: An algorithm for nudity detection with KNN classifier [43].
Algorithm F: An algorithm for nudity detection with SVM classifier [43].
Algorithm G: A practical system for detecting obscene videos [15].

Table 6 illustrates the average RR for the implemented algorithms. We tested the algorithms with the same data as the proposed algorithm. We examined KNN classifier with various K values, and SVM classifier with different kernels, and the best results are reported in Table 6. As shown in this table, the maximum recognition belongs

Table 5 Average RR for the proposed algorithm when various features elements are removed from the proposed feature vector

Removed features	SVM core	Average RR (%)
Features based on the information of single frames	RBF	75.8
	Linear	84.5
	Quadratic	80.6
	Polynomial	83.2
Features based on 3D STV (first group)	RBF	77.3
	Linear	92.1
	Quadratic	84.8
	Polynomial	83.5
Features based on 3D STV (second group)	RBF	77.1
	Linear	93.7
	Quadratic	85.3
	Polynomial	80.6
Features based on motion and periodicity characteristics	RBF	74.5
	Linear	94.7
	Quadratic	89
	Polynomial	86.7

Table 7 RR of the proposed algorithm for various categories of the collected database

Categories	RR						
	Exp. #1 (%)	Exp. #2 (%)	Exp. #3 (%)	Exp. #4 (%)	Exp. #5 (%)	Average (%)	Std. deviation (%)
Animal sex	63.20	72.10	78.23	68.30	69.00	70.17	5.53
Bad illumination	77.16	83.71	86.90	84.32	74.72	81.36	5.16
Porn with cloth	57.45	58.33	67.73	64.50	65.35	62.67	4.53
Animation (porn and non-porn)	94.56	96.43	97.50	97.13	96.23	96.37	1.14
Low-resolution video clips	68	85	87.92	79.30	76.44	79.39	7.85
Heterosexual	96.40	97.90	98.68	94.20	98.90	97.22	1.95

Table 8 The result of the proposed algorithm for the recognition of animal and animation obscene videos after retraining each category individually

Categories	RR						
	Exp. #1 (%)	Exp. #2 (%)	Exp. #3 (%)	Exp. #4 (%)	Exp. #5 (%)	Average (%)	Std. deviation (%)
Animal sex	92.47	96.77	95.69	97.84	98.92	96.34	2.47
Animation	96.78	97.63	92.28	97.49	98.30	96.50	2.42

to the Algorithm C. Comparisons between the results of Tables 2 and 6 show that the proposed algorithm has improved the RR by 9.34%.

4.1. Error analysis

To analyze error sources for the proposed algorithm, we tested the proposed algorithm with different categories of database videos. Table 7 shows the RR of the proposed algorithm for various categories of the collected database. As it is obvious from the table, the algorithm has lower RR for some categories, including animal sex, bad illumination, porn with clothes, and low-resolution video clips. Some reasons for the error source of these categories are

- skin detection algorithm may fail in some video episodes, especially in low-resolution videos or frames with bad illumination;
- there may be no considerable skin region in the frames;
- for some animal sex videos, feature vector elements are slightly different from traditional obscene videos.

In other experiments, we retrained an SVM classifier for the recognition of obscene animal videos. In these experiments, only animal sex videos were used as obscene videos. The same experiments were repeated for obscene animation videos recognition as well. Table 8 illustrates the results of these experiments. The results show that when each category is trained individually the SVM classifier shows better discrimination. Therefore, as a future work we are going to use combined classifiers for further improving the RR.

5. Conclusions

In this article, a new method for the recognition of obscene video contents was presented. We used SVM classifier with three novel sets of features for the recognition of video episodes. The proposed features were based on spatial, ST, and periodicity characteristics of skin regions in the video episodes. In order to evaluate the proposed algorithm, a database of video files was collected by random web surfing. The proposed algorithm was implemented using Microsoft Visual C++ compiler by using DirectX SDK facilities. Experimental results showed the RR of 95.44% for the proposed algorithm. We compared the results of the proposed algorithm with those of other methods, and the results showed that the proposed algorithm improves the RR by 9.34%. As a future work, we plan to use combined classifiers for further improving the RR.

Abbreviations

MID: Motion information difference; PCA: Principal component analysis; RR: Recognition rate; STV: Spatiotemporal volume; SVM: Support vector machine; XMAS: X multimedia analysis system.

Competing interests

The authors declare that they have no competing interests.

Acknowledgment

This study was supported by the Iranian Research Institute for ICT (ITRC).

Author details

[1]Faculty of Engineering, Shahed University, Tehran, Iran. [2]Iranian Research Institute for ICT (ITRC), Tehran, Iran.

Received: 4 May 2012 Accepted: 20 November 2012
Published: 19 December 2012

References

1. D.A. Forsyth, M.M. Fleck, Automatic detection of human nudes. Int. J. Comput. Vis. **32**(1), 63–77 (1999)
2. J. Yang, Z. Fu, T. Tan, W. Hu, A novel approach to detecting adult images, in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 4, England, UK*, 2004, pp. 479–482
3. J. Ze Wang, J. Li, G. Wiederhold, O. Firschein, System for screening objectionable images. Comput. Commun. **21**(15), 1355–1360 (1998)

4. R. Du, R. Safavi-Naini, W. Susilo, Web filtering using text classification, in *Proceedings of the 11th IEEE International Conference on Networks (ICON2003), Sydney, NSW, Australia*, 2003, pp. 325–330
5. Q. Wang, W.M. Hu, T.N. Tan, Detecting objectionable videos. Acta Automatica Sinica **31**(2), 280–286 (2005)
6. B. Choi, J. Kim, J. Ryou, Retrieval of illegal and objectionable multimedia, in *Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management (NCM '08), Gyeongju, Korea*, 2008, pp. 645–647
7. C.Y. Kim, O.J. Kwon, W.G. Kim, S.R. Choi, Automatic system for filtering obscene video, in *Proceedings of the 10th International Conference on Advanced Communication Technology (ICACT 2008), Gangwon-Do, South Korea*, vol. 2, 2008, pp. 1435–1438
8. N. Rea, G. Lacey, C. Lambe, R. Dahyot, Multimodal periodicity analysis for illicit content detection in videos, in *Proceedings of the 3rd European Conference on Visual Media Production, London, UK*, 2006, pp. 106–114
9. Q. Zhiyi, L. Yanmin, L. Ying, J. Kang, C. Yong, A method for reciprocating motion detection in porn video based on motion features, in *Proceedings of the 2nd IEEE International Conference on Broadband Network & Multimedia Technology (IC-BNMT '09), Beijing, China*, 2009, pp. 183–187
10. C. Jansohn, A. Ulges, T.M. Breuel, Detecting pornographic video content by combining image features with motion information, in *Proceedings of the 17th ACM international conference on Multimedia, New York, NY, USA*, 2009, pp. 601–604
11. H. Lee, S. Lee, T. Nam, Implementation of high performance objectionable video classification system, in *Proceedings of the 8th International Conference on Advanced Communication Technology (ICACT 2006), vol. 2, Phoenix Park, Gangwon-Do, Korea*, 2006, pp. 959–962
12. S. Lee, W. Shim, S. Kim, Hierarchical system for objectionable video detection. IEEE Trans. Consum. Electron. **55**(2), 677–684 (2009)
13. S. Zhao, L. Zhuo, S. Wang, L. Shen, Research on key technologies of pornographic image/video recognition in compressed domain. J. Electron. (China) **26**(5), 687–691 (2009). doi:10.1007/s11767-009-0020-8
14. A.P.B. Lopes, S.E.F. de Avila, A.N.A. Peixoto, R.S. Oliveira, Nude detection in video using bag-of-visual-features, in *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), Rio de Janeiro, Brazil*, 2009, pp. 224–231
15. C.Y. Kim, O.J. Kwon, S. Choi, A practical system for detecting obscene videos. IEEE Trans. Consum. Electron. **57**(2), 646–650 (2011)
16. A. Akbulut, F. Patlar, C. Bayrak, E. Mendi, J. Hanna, Agent based pornography filtering system, in *International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Trabzon, Turkey*, 2012, pp. 1–5
17. P.M. da Silva Eleuterio, M. de Castro Polastro, B.F. Police, An adaptive sampling strategy for automatic detection of child pornographic videos, in *Proceedings of the Seventh International Conference on Forensic Computer Science, Brasilia, DF, Brazil*, 2012, pp. 12–19
18. H.J. Zhang, J. Wu, D. Zhong, S.W. Smoliar, An integrated system for content-based video retrieval and browsing. Pattern Recognit. **30**(4), 643–658 (1997)
19. W. Wolf, Key frame selection by motion analysis, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, USA*, 1996, pp. 1228–1231
20. L. Li, X. Zhang, Y. Wang, W. Hu, P. Zhu, Nonparametric motion feature for key frame extraction in sports video, in *Proceedings of the Chinese Conference on Pattern Recognition (CCPR '08), Beijing, China*, 2008, pp. 1–5
21. C. Kim, J.N. Hwang, An integrated scheme for object-based video abstraction, in *Proceedings of Eighth ACM international conference on Multimedia, New York, NY, USA*, 2000, pp. 303–311
22. X. Song, G. Fan, Joint key-frame extraction and object segmentation for content-based video analysis. IEEE Trans. Circuits Syst. Video Technol. **16**(7), 904–914 (2006)
23. J. Jiang, X.P. Zhang, Gaussian mixture vector quantization-based video summarization using independent component analysis, in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP), Saint Malo, France*, 2010, pp. 443–448
24. L. Zhao, W. Qi, S.Z. Li, S.Q. Yang, H. Zhang, Key-frame extraction and shot retrieval using nearest feature line (NFL), in *Proceedings of the ACM Multimedia Workshop 2000, Los Angeles, CA*, 2000, pp. 217–220
25. C.I. Chang, Y. Du, J. Wang, S.M. Guo, P. Thouin, Survey and comparative analysis of entropy and relative entropy thresholding techniques. IEE Proc. Vis. Image Signal Process. **153**(6), 837–850 (2006)
26. M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection. Int. J. Comput. Vis. **46**(1), 81–96 (2002)
27. Y. Wang, B. Yuan, A novel approach for human face detection from color images under complex background. Pattern Recognit. **34**(10), 1983–1992 (2001)
28. F. Chang, Z. Ma, W. Tian, A region-based skin color detection algorithm. Adv. Knowledge Discover. Data Min. **4426**, 417–424 (2007)
29. P. Ruangyam, N. Covavisaruch, An efficient region-based skin color model for reliable face localization, in *Proceedings of the 24th International Conference Image and Vision Computing New Zealand (IVCNZ '09), Wellington, New Zealand*, 2009, pp. 260–265
30. H. Bouirouga, S. El Fkihi, A. Jilbab, M. Bakrim, A comparison of skin detection techniques for objectionable videos, in *Proceedings of the 5th International Symposium on I/V Communications and Mobile Network (ISVC), Rabat, Morocco*, 2010, pp. 1–4
31. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press Professional, New York, 1990)
32. C.W. Su, H.Y.M. Liao, H.R. Tyan, C.W. Lin, D.Y. Chen, K.C. Fan, Motion flow-based video retrieval. IEEE Trans. Multimed. **9**(6), 1193–1201 (2007)
33. Y. Du, F. Chen, W. Xu, W. Zhang, Activity recognition through multi-scale motion detail analysis. Neurocomputing **71**(16–18), 3561–3574 (2008)
34. T.H.W. Lam, R.S.T. Lee, Human identification by using the motion and static characteristic of gait. Pattern Recognit. **3**, 996–999 (2006)
35. M. Allmen, C.R. Dyer, Cyclic motion detection using spatiotemporal surfaces and curves, in *Proceedings of the 10th International Conference on Pattern Recognition, vol. 1, Atlantic City, NJ, USA*, 1989, pp. 365–370
36. A. Briassouli, N. Ahuja, Extraction and analysis of multiple periodic motions in video sequences. IEEE Trans. Pattern Anal. Mach. Intell. **29**(7), 1244–1261 (2007)
37. F. Cheng, W. Christmas, J. Kittler, Periodic human motion description for sports video databases, in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 3, Cambridge, England, UK*, 2004, pp. 870–873
38. R. Cutler, L. Davis, View-based detection and analysis of periodic motion, in *Proceedings of the Fourteenth International Conference on Pattern Recognition, vol. 1, Brisbane, QLD, Australia*, 1998, pp. 495–500
39. X. Tong, L. Duan, C. Xu, Q. Tian, H. Lu, J. Wang, J.S. Jin, Periodicity detection of local motion, in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2005), Amsterdam, Netherlands*, 2005, pp. 650–653
40. V.N. Vapnik, *Statistical Learning Theory* (Wiley-Interscience, 1998)
41. C.J.C. Burges, *Advances in Kernel Methods: Support Vector Learning* (The MIT Press, Cambridge, MA, 1999)
42. H. Zheng, M. Daoudi, B. Jedynak, Adult image filtering for web safety, in *Proceedings of the 2nd International Symposium on Image/Video Communications over Fixed and Mobile Networks, Brest, France*, 2004, pp. 77–80. Adult image filtering for web safety
43. R. Ap-apid, An algorithm for nudity detection, in *Proceedings of the 5th Philippine Computing Science Congress, Cebu City, Philippines*, 2005, pp. 199–204

doi:10.1186/1687-5281-2012-23
Cite this article as: Behrad *et al.*: Content-based obscene video recognition by combining 3D spatiotemporal and motion-based features. *EURASIP Journal on Image and Video Processing* 2012 **2012**:23.