Kullback-Leibler similarity measures for effective content based video retrieval

R Priya*[a], **T N Shanmugam**[a] and **R Bhaskaran**[b]

[a]Department of Mathematics, Anna University, Chennai, Tamil Nadu, India

[b]Department of Computer Science and Engineering, Anna University, Chennai, Tamil Nadu, India

Abstract: Recent advancements in the multimedia technologies allow the capture and storage of video data with relatively inexpensive computers. As the necessity to query these data competently becomes significant, the amount of broadly accessible video data grows. As a result, content-based retrieval of video data turns out to be a demanding and vital problem. In this paper, an effective content-based video retrieval system is proposed. The raw video data are segmented into shots and the object feature, movement feature and the occlusion feature are extracted from these shots and the feature library is utilised for the storage process of these features. Subsequently, the Kullback–Leibler distance is computed among the features of the feature library and the features of the query clip which is extracted in the similar manner. Hence, with the aid of the Kullback–Leibler distance, the similar videos are extracted from the collection of videos based on the given query video clip in an effective manner.

Keywords: video retrieval, content-based video retrieval (CBVR), video sequence, shot segmentation, object feature, movement feature, occlusion feature, query clip, similarity measure, Kullback–Leibler distance

1 INTRODUCTION

Video retrieval is a vital technology which is used to design video search engines which act as an information filter and an initial set of relevant videos is sifted out from the database.[1] By matching those videos in the database that are nearest to the query object in high dimensional spaces with the feature attributes of the query object, the process of retrieval is performed.[2] The retrieval of video data which is based on their visual content like colour distribution, texture and shape and which works in accord with similarity measurement has been the main focus of a lot of researchers.[4]

Low-level visual features and high-level semantic content are the two stages in which the video content can be grouped.[10] Colour, texture and shape are the extraction of low-level features.[13] Semantic content extraction is more complex because it is not only based on low-level features (colour, texture, shape, object, etc.), i.e. visual similarity, but also necessitates domain knowledge, user interaction and semantic extraction.[12]

On the basis of the video contents through user interactions, content-based video retrieval (CBVR) can competently aid users to retrieve preferred video segments from a large video database.[6] A very frequent first step in the majority of content-based video[35] analysis techniques is to segment a video into elementary shots, each consisting of a continuous time and space.[3] A series of frames with uninterrupted camera motion is known as a shot, while a

The MS was accepted for publication on 19 April 2012.

* *Corresponding author: R Priya, Department of Mathematics, Anna University, Chennai, Tamil Nadu, India; email: priyarajendranphd@gmail.com*

IMAG 306 © RPS 2013 DOI: 10.1179/1743131X12Y.0000000017

series of shots that are consistent from the narrative as well as the users' point of view is known as a clip.[19]

Video segmentation is a major manoeuver in image sequence analysis and its results are widely employed for determining motion features of scene objects, as well as for coding functions to decrease storage requirements.[7] Video segmentation splits a video file into shots which is illustrated as a contiguous sequence of video frames that are recorded from a single camera operation.[11] Edge information-based video segmentation, image segmentation-based video segmentation and change detection-based video segmentation are the three types of video segmentation.[8] The feature extractor, which is a real-time system, is employed to pre-process all the videos that are stored in the database and also store their unique features for quicker retrieval.[9] For locating contents similar to a query video stream from video database, feature matching is employed.[15] To represent appropriate objects which appear in those shots, the shot features are grouped into clusters in the grouping stage.[5]

Frame sequence matching and key-frame-based shot matching are the two categories in which CBVR[34] is classified.[14] Frame-based approaches presume a set of key frames that can offer a compact representation of commercial video contents.[18] A frame that can represent the significant content of a video shot is known as a video key frame.[16] The key frames are directly extracted from the video sequence; therefore the added computational overhead is not necessary when compared with shot-based key frame extraction techniques.[20] Key frame extraction allows fast video surfing and it also offers a powerful tool for video content summarisation and visualisation.[17]

Generally, movement feature and object feature play a very vital role in retrieving the similar video clips. However, to improve the performance of the retrieval, the features need to be enhanced. The existing techniques extract the features by different means. In this paper, we enhance the (1) object feature by extracting the spatial information of the detected object and (2) movement feature by recognising the direction of movement of the feature of the subjected video clip. Additionally, we extract the occlusion feature for further improvement in the performance of the system.

As a pre-processing stage, the raw video data are segmented into frames. After that the frames are grouped to their corresponding shots for the consecutive processes. Subsequently, for the retrieval of the video, based on the given query video, the features are extracted based on the object, movement and occlusion. In order to obtain the number of objects presented in the shots, the fuzzy K-means clustering is utilised and to identify the location of the objects, the spatial feature is also combined with it and then the movement feature is extracted from the video.

As briefed earlier, the movement feature is mainly concentrated on the posture and movement of the humans presented in the video, the direction and distance are identified. Subsequently, the occlusion feature is identified and then all these features of a video are combined as a feature set in the feature library. In the proposed system, the videos are retrieved based on the query clip. For the query video clip, the aforesaid object, movement and occlusion features are extracted and compared with the feature in the feature library. The comparison is achieved via the Kullback–Leibler distance (KLD) similarity measure. Afterwards, the similar videos are retrieved from the collection of videos.

The rest of the paper is organised as follows. A short review of a few of the existing works in content-based video retrieval is presented in Section 2. The proposed effective CBVR system is detailed in Section 3. The results and discussion are described in Section 4. The conclusions are summed up in Section 5.

2 REVIEW ON RELATED RESEARCHES

A motion-based video retrieval technique was proposed by Hsieh *et al.*[21] for retrieving required video sequences as per their trajectory features. Sketch-based and string-based techniques are the two complementary techniques that were incorporated in this method for representing and indexing trajectories with more syntactic meanings. Thus, diverse trajectories can not only be compared from their low-level features but also their syntactic meanings. By the proposed technique, major developments in video accuracy have been achieved, since most impossible candidates can be filtered out by employing syntactic meanings. Moreover, the problem of partial trajectory matching has been easily solved by this technique.

A technique based on spatiotemporal independent component analysis (stICA) and multiscale analysis used to extract objects from video sequences has been proposed by Zhang and Chen.[22] The preliminary source images which comprise the moving objects in video sequences can be extracted by stICA. To enhance the precision of the removed object, the wavelet-based multiscale image segmentation and region detection techniques process the source image data attained after stICA analysis. On the basis of those projected techniques, an automated video object extraction system is developed. In the content-based video processing applications, the preliminary outcome illustrates huge prospective for the projected stICA and multiscale-segmentation-based object extraction system.

The use of motion information in the compressed domain allows the rapid analysis of the content of the video and this was presented by Babu and Ramakrishnan.[23] On the basis of the motion information attained from the compressed MPEG video, a video indexing and retrieval system was presented. Employing the projected K-means algorithm on the refined motion data determines the number of objects in the video shot and by segmenting the objects by expectation–maximisation (EM) algorithm, the object features are attained. For the process of retrieval, the global and object features with the user given weights were employed. By taking into account the spatial features such as DCT dc coefficients that can be easily extracted from the MPEG video, the system can be further enhanced.

The technology of moving-object tracking to content-based video retrieval was proposed by Wen *et al.*[24] Background subtraction was employed to detect moving pixels, and overcome the shadow problem, and then connected components labelling and morphological operations were employed to eradicate noise and fix the moving pixels. Then, with colour histograms, colour similarity and 'motion vector', the target's image and the information for content-based video retrieval in the database can be extracted. For image frame retrieval in single charge coupled device (CCD) or multi-CCD surveillance systems, this technique can be implied. Detection and retrieval error for multi-surveillance retrieval can be caused by rapid environment changes (such as light), CCD shift, viewing angle and position.

By analysing the characteristics of spatiotemporal volumes in videos, a strong video matching framework was proposed by Basharat *et al.*[25] On the basis of the clustering of the interest point trajectories, the volumes were built. Multiple features, namely colour, texture, motion and interest point descriptors were extracted to model the appearance of the volumes. By resolving the utmost matching problem of the graph that was created by the volumes, the likeness between two videos was computed. A very capable and aggressive performance in video matching for retrieval was achieved by utilising the proposed video matching framework.

By merging the application and network level parameters for all content types, the prediction of video quality was proposed by Khan *et al.*[26] First, by employing cluster analysis, the video sequences were classified into groups that represent different content types. The classification of contents was on the basis of the temporal (movement) and spatial (edges and brightness) feature extraction. Second, the behaviour of video quality is studied and analysed for wide range variations of a set of chosen parameters. Finally, two learning models which are based on (1) ANFIS to guess the visual perceptual quality in terms of the mean opinion score and decodable frame rate (Q value) and (2) regression modelling to guess the visual perceptual quality in terms of the mean opinion score were developed.

To detect and extract the text from the video scene, a framework was proposed by Pratheeba *et al.*[27] By computing the disparity between the closing and the opening images, a morphological binary map was engendered. Then by employing a morphological dilation operation, the candidate regions were connected and based on the incidence of text in every candidate, the text regions are resolved. By employing the projection of text pixels in the morphological binary map, the detected text regions were precisely localised and finally the text extraction was conducted. The projected technique was strong to diverse character size, position, contrast and colour. It was also language-independent. To lessen the processing time, text region update between the frames was deployed.

In our previous research,[30] we have proposed an adaptive system with broad features to improve the efficiency of the retrieval system. Our proposed system segmented a video into shots, and then a few representative frames were generated from each of the shot on the basis of different features such as colour, contour, texture and motion, and these frame descriptors were calculated for these shots and were stored in a feature library. When a query for the clip has been given, the features mentioned above were extracted for the query clip and then were compared with the features that are already stored in the feature library. A technique called latent semantic indexing on the basis of similarity measure was used to perform the comparison. Finally, videos similar to the query are obtained from the vast collection of videos. Our system has been evaluated with precision-recall and F score technique and we have compared with existing retrieval systems.

3 PROPOSED CONTENT BASED VIDEO RETRIEVAL SYSTEM

This section mainly deals with describing the proposed video retrieval system. The system mainly consists of the following sections.

1. Shot segmentation.
2. Object-based feature extraction.
3. Movement-based feature extraction.
4. Occlusion feature extraction.
5. Retrieving relevant video clips.

They are detailed further. Let V_i, $i=1,2, \ldots, N_v$ be a database video, where N_v is the total number of videos present in the database. The video is a collection of frames with size $M \times N$ that can be represented by $f_j^{(i)}(x,y)$, $j=1, 2, \ldots, N_f^{(i)}$; $x=0, 1, 2, \ldots, M-1$ and $y=0, 1, 2, \ldots, N-1$, where $N_f^{(i)}$ is the total number of frames present in the ith database video.

3.1 Shot segmentation

A video is a collection of a huge number of still frames and as the continuous change of the frames exhibits a motion like feature, it is known as video. However, the video has different shots, which can be defined as a sequence of frames taken by a single camera without any major variation in the colour content of consecutive videos. To process any video, or to extract the features from the video, it is necessary to segment the video by means of different shots. As stated earlier, the database video V_i has $N_f^{(i)}$ frames and the frames are grouped based on the shots. To accomplish this, initially each frame is split into different blocks of size $m \times n$ and DCT is applied to every block of the frame as follows

$$B_{j,k}^{(i)}(c,d)= \left(\frac{2}{m}\right)^{1/2}\left(\frac{2}{n}\right)^{1/2} \alpha_c \alpha_d \sum_{a=0}^{m-1}\sum_{b=0}^{n-1} b_{j,k}^{(i)}(a,b) \cos\left[\frac{\pi c}{2m}(2a+1)\right]\cos\left[\frac{\pi d}{2n}(2b+1)\right] \tag{1}$$

where

$$\alpha_c = \begin{cases} \left(\frac{1}{2}\right)^{1/2} & \text{if} c=0 \\ 1\text{if} & 1 \le c \le m-1 \end{cases} \tag{2}$$

$$\alpha_d = \begin{cases} \left(\frac{1}{2}\right)^{1/2} & \text{if } d=0 \\ 1\text{if} & 1 \le d \le n-1 \end{cases} \tag{3}$$

In equation (1), $k=1, 2, \cdots, N_{b_j}^{(i)}$, where $N_{b_j}^{(i)}$ is the number of blocks present in the jth frame of the ith video clip. Once the blocks of every frame are converted to transform domain, Euclidean distance is determined for the blocks of every consequent frame as follows

$$E_{d_j}^{(i)} = \frac{1}{N_{b_j}^{(i)}} \sum_{k=1}^{N_{b_j}^{(i)}} \left\{ \sum_{c=0}^{m-1}\sum_{d=0}^{n-1} \left[B_{j,k}^{(i)}(c,d) - B_{j+1,k}^{(i)}(c,d) \right]^2 \right\}^{1/2} \tag{4}$$

Based on the determined Euclidean distance, the frames that belong to the same shots are determined. This can be achieved using some criterions as follows: (1) if $E_{d_j}^{(i)} \le E_T, f_j^{(i)}$ and $f_{j+1}^{(i)}$ belong to the same shot and (2) if $E_{d_j}^{(i)} > E_T, f_j^{(i)}$ and $f_{j+1}^{(i)}$ belong to different shots. Hence, different shots are extracted from the subjected video and so $N_s^{(i)}$ shots are obtained from every ith video clip. The obtained shots are also a collection of frames and on the basis of shot, the further process of feature extraction is performed.

3.2 Object-based feature extraction

In Ref. 5, an object-based video retrieval system has been proposed. The work involves extraction of object feature from the video sequence. However, it shows inaccuracy because it does not consider the spatial feature, i.e. the location of the object is not taken into account. The proposed system considers the spatial feature and overcomes the aforesaid practical issues. In order to extract the object feature, the frames are identified and assembled to their corresponding shots s_a, $a=0, 1, 2, \ldots, N_s^{(i)}$ and the number of frames in a shot s_a is determined as $|s_a|$. The shot s_a contains $|s_a|$ frames; the initial frame $f_0^{(i)}$ is assumed as the key frame $f_{key}^{(i)}$ and each shot s_a has its own key frame. Initially, the frames $f_j^{(i)}$ of shot s_a which are in the RGB colour are converted to the grey scale component

$$f_{jgv}^{(i)} = 0{\cdot}2989 \times f_{jr}^{(i)} + 0{\cdot}5870 \times f_{jg}^{(i)} + 0{\cdot}1140 \times f_{jb}^{(i)} \tag{5}$$

The above equation is the Craig's formula for converting RGB colour video to grey scale video.

Then the clustering process is applied to the converted grey scale key frame images for object identification. The video is dynamic, so a static K value cannot be obtained. Hence, the number of objects for the clustering process is determined using the 3D colour histogram. After converting the RGB colour space $f^{(i)}_{key}$ to the LAB colour space, the 3D colour histogram is determined for the key frame $f^{(i)}_{key}$. From the obtained 3D colour histogram of the selected key frame $f^{(i)}_{key}$, the number of peaks, say K, which is also the number of objects is identified. Thus, K clusters are obtained by the clustering process. Fuzzy K-means clustering facilitates the identification of overlapping groups of objects because it allows the objects to have membership in more than one group. The pseudo code for the fuzzy K-means clustering is shown below

```
    Input: grey scale converted frames $f^{(i)}_{jgv}$ in a shot $s_a$
    $K-$number of clusters
    Output: set of $K$ clusters
• For each $f^{(i)}_{jgv}$ to $f^{(i)}_{jgv_{|s_a|}}$
• Arbitrarily select $K$ pixels $P_1, P_2, \ldots, P_K$ from $f^{(i)}_{jgv}$
  as initial centroids.
• Until there are no changes in any mean
• Use the estimated means to find the degree of
  membership $\mu(y,x)$ of $f^{(i)}_{jgv}$ in cluster $\eta$
• For $\eta$ from 1 to $K$
  Replace $P_\eta$ with the fuzzy mean of all of the
  examples for cluster $\eta$
      $P_\eta = \frac{\sum_{y} \mu(y,x)^2 f^{(i)}_{jgv}}{\sum_{y} \mu(y,x)^2}$
• End for
End until
    End for
```

Hence, K clusters are identified from the grey scale converted frames of the shot s_a.

3.2.1 Track frame selection

The extracted clusters of the key frame $f^{(i)}_{key}$ are the objects and they are compared with the other frames of the shot for their presence in the frame. If the object is presented in the remaining frames in the shot, then the corresponding frame index is stored in the $TF^{(i)}_{id}$. This vector contains the index of the track frame index which is having at least a single object. This track frame selection is utilised to reduce the computational time because a frame is said to be a track frame where at least a single object is presented in that particular frame. Hence in order to extract the object-based features, rather than analysing all the frames, these track frames are analysed. The following pseudo code details the selection of track frames.

```
For each $f^{(i)}_{j_{gv}}$ to $f^{(i)}_{j_{gv_{|s_a|}}}$
  For $\forall P'_\eta$,  $\eta=1, 2, \ldots, K$
    For $\forall P_\eta$, $\eta=1, 2, \ldots, K$
      $L2_{pq}=\left\{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\left[\left|P'_\eta(x,y)-P_\eta(x,y)\right|^2\right]\right\}^{1/2}$
      If $L2_{pq}\leqslant 0.6$ then
        $TF^{(i)}_{id}=j$
      End if
    End for
  End for
End for
```

3.2.2 Cluster grouping

After the selection of track frames, the clusters are grouped by analysing the track frames and then these clusters presented in the remaining frames append to their corresponding clusters. Hence, the K clusters are grouped and the cluster group is used for the further process. The pseudo code shown below details the process.

```
For each $f_{TF^{(i)}_{id}}$,  $id=1, 2, \ldots, N_{TF^{(i)}_{id}}$d
  For $\forall P'_\eta$,  $\eta=1, 2, \ldots, K$
    For $\forall P_\eta$, $\eta=1, 2, \ldots, K$
      $L2_{pq}=\left\{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\left[\left|P'_\eta(x,y)-P_\eta(x,y)\right|^2\right]\right\}^{1/2}$
      $G^{(i)}_\eta = P'_\eta(x,y)$
      If $L2_{pq}\leqslant 0.6$ then
        $G^{(i)}_\eta \cup P_\eta(x,y)$
      End if
    End for
  End for
End for
```

Then the cluster group $G_{\eta}^{(i)}(x,y)$ for the shot s_a and the covariance for this cluster group are computed

$$C_{s_a}^{(i)}(x,y)=\sum\{[x-E(x)][y-E(y)]\}/(n-1) \qquad (6)$$

where $E(x)$ is the mean of x and $E(y)$ is the mean of the y values. Let $Ob_{s_a}^{(i)}(w)$, hence the object feature is extracted and the drawback in this feature is that it cannot predict the location of the object wherever the object is identified that it marks as a feature. Hence the spatial feature is also extracted. The object feature and the spatial feature are combined and used as the feature set for the further process.

3.2.3 Spatial feature extraction

The spatial feature is extracted to identify the location of the object. The following pseudo code details the process of obtaining the appropriate objects to be identified as the spatial information.

```
For each $TF_{id}^{(i)}$,  $id=1, 2, \ldots, N_{TF_{id}}^{(i)}$
  For each $Ob_{TF_{id}}^{(i)}(w)$,  $w=1, 2, \ldots, N_w$
    If size$\left[Ob_{key}^{(i)}(w)\right]>$size$\left[Ob_{TF_{id}}^{(i)}(w)\right]$
    Resample size of $Ob_{key}^{(i)}(w)$
    Else
    Resample size of $Ob_{TF_{id}}^{(i)}(w)$
    End if
$D_{(w)}^{(i)}=Ob_{key}^{(i)}(w)-Ob_{TF_{id}}^{(i)}(w)$
Normalize $D_{(w)}^{(i)}$ and select minimum $D_{(w)}^{(i)}$
  End for
End for
```

In order to accomplish the spatial information, the clustered key frame $f_{key}^{(i)}(x,y)$ is re-sampled and then $M\times N$ frame is created with each pixel of 0 values ($M=N=128$). After that the frame is alienated to blocks of $M/4\times N/4$. Each object $D_{(w)}^{(i)}$ is placed on the empty white pixelled frame. The number of 1's in each block is analysed and then the index of the block, which is of maximum number of 1's, is stored in the vector $sp_{s_a}^{(i)}(K)$. Hence the object feature is also combined with the spatial feature and utilised for the subsequent processes. $C_{s_a}^{(i)}(x,y)$ and $sp_{s_a}^{(i)}(K)$ are the extracted object features.

3.3 Movement-based feature extraction

An enhanced object's movement classification system is proposed to extract the object's movement-based feature from the video sequences. In Ref. 28, a posture classification system has been proposed for the same. The work considers the posture and movement of only humans present in the video sequence generally. However, it shows inaccuracy because it does not considers the location of movement, i.e. whether the human is moving in front of the screen or away from the screen and direction of movement, whether the movement is from left to right or from right to left as well as whether the movement is towards the screen or away from the screen. The aforesaid practical issues are considered and they are overcome in the proposed system. The proposed system extracts the features such as skeleton, centroid context, object orientation and direction of movement from a foreground object. Prior to extracting the features, the video shots are subjected to foreground segmentation using minimum graph cut algorithm.[29] The foreground segmented frames are subjected to extraction of the following features.

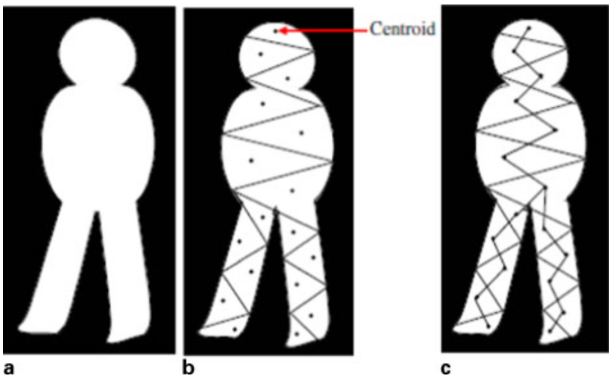
3.3.1 Extraction of centroid feature

All the obtained objects are subjected to triangulation so that the objects are filled up with networked triangles. For each triangle, a centroid is determined. The triangulated objects with centroids are given in Fig. 1.

Hence for every frame, the foreground objects are determined and the centroids are determined. The coordinates of the obtained centroids are stored as centroid features for the particular video clip.

3.3.2 Extraction of orientation and distance of object movement feature

The distance of object movement is defined here that at which distance from the screen, the object exhibits movement in the video whereas orientation is defined here that the type of movement of the object (very specifically humans). In order to accomplish this, the skeleton of the object is obtained from Fig. 1. With



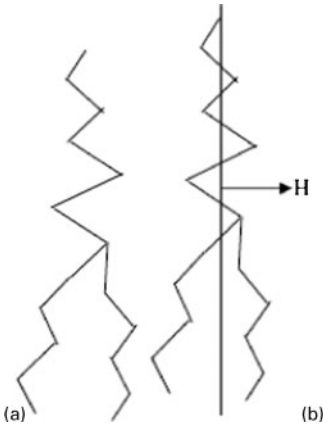
1 Extraction of centroid feature: (a) foreground segmented frame; (b) triangulised and centroid marked frame; (c) skeletonised frame

the reference of the skeleton, the axis of the object is determined. The obtained skeleton and the axis formation are depicted in Fig. 2. Trigon is formed by resizing the frames into dissimilar sizes and processed as per the size of the frame.

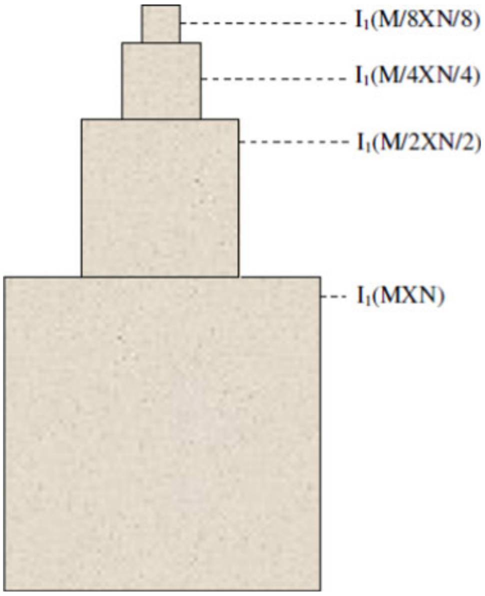
Figures 20 and 21 show the combined result for the object feature and the spatial feature indicates that similar frames of the input video are retrieved based on the given query video frame and compared with object feature-based extraction, combined feature-based extraction has produced better results.

From the obtained axis, i.e. altitude of the object, the decision is made. When the axis satisfies the condition $H \geqslant H_{TH}$, where H is the axis, then it is decided that the object is moving in front of the screen. Otherwise, the object is moving away from the screen at a greater distance. The orientation is determined by calculating the angle between the axis of the object and ground plane.

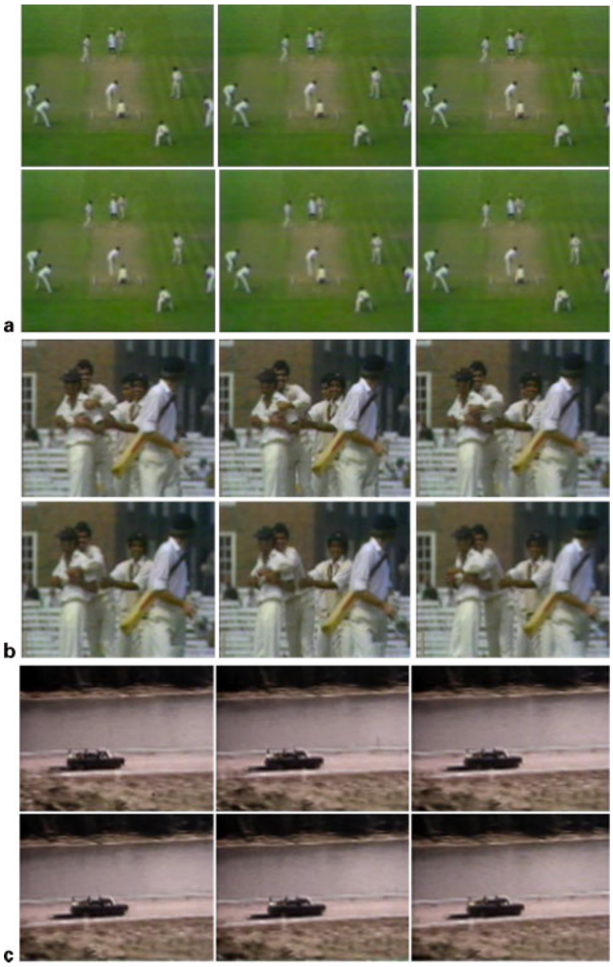
In Fig. 5 RGB image and its corresponding histogram and 3D image are showed and in Fig 6



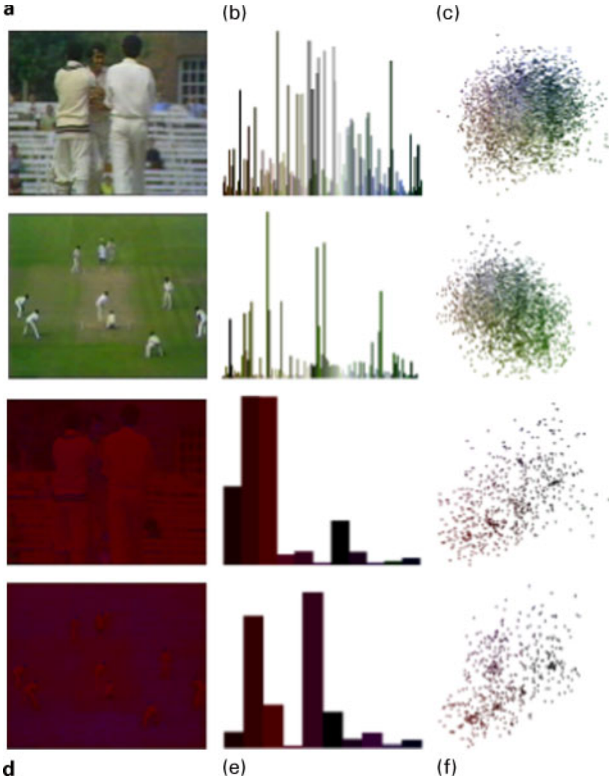
2 Extraction of distance of movement: (a) skeleton of the object; (b) axis formation



3 Trigon formation



4 (a–c) Different shots of elementary frames are segmented amid dissimilar video sequences

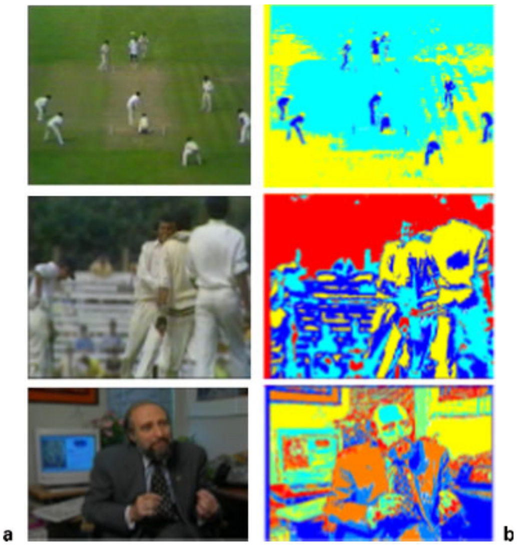


5 (a) Keyframe in RGB colourspace, and their corresponding (b) histogram and (c) 3D colour histogram, and (d) LAB colourspace of keyframe, and their corresponding (e) histogram and (f) 3D colour histogram.

shows the Key frames and their corresponding and its corresponding clustered frames.

3.3.3 *Extraction of direction of movement*

The foreground segmented frames are considered to extract the direction of movement of the frames. The segmented frames are subjected to derive the direction of movement. The direction of movement is obtained by analysing all the frames present in a shot in four scenarios. Prior to executing the scenarios, row and column origins O_{ab}^{row} and O_{ab}^{col}, respectively are determined as follows



6 (a) Keyframes and their corresponding (b) clustered frames

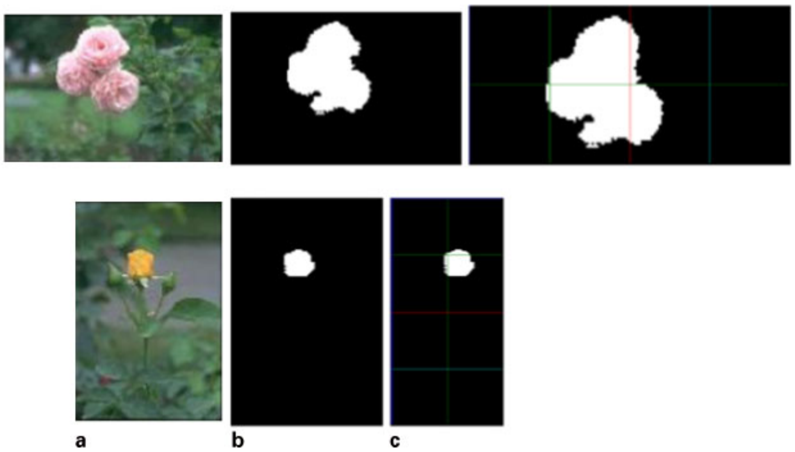
$$O_{ab}^{row}, \quad \text{if} \quad F_{ab}(j,i)=1 \tag{7}$$

$$O_{ab}^{col}, \quad \text{if} \quad F_{ab}(i,j)=1 \tag{8}$$

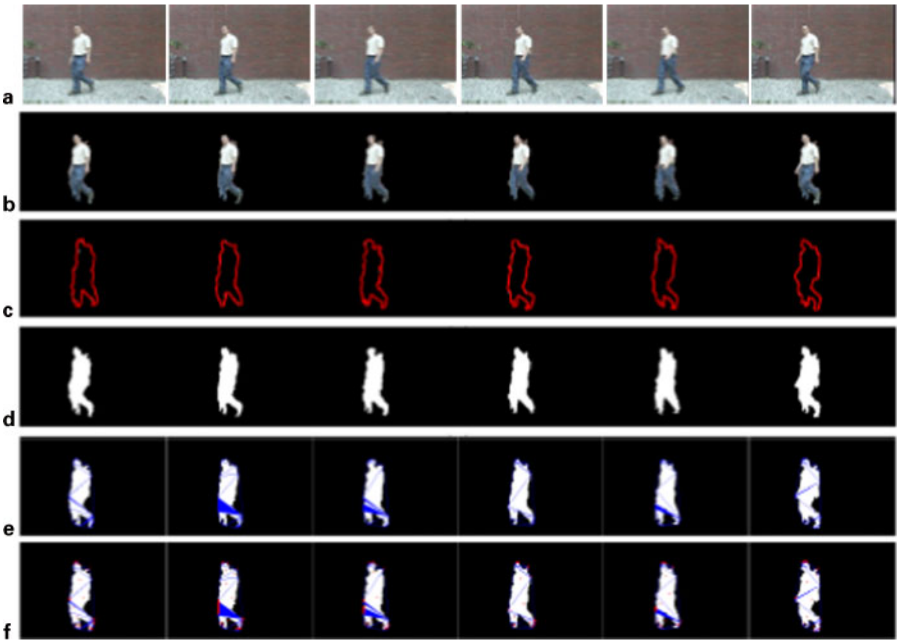
where $a=0, 1, 2, \ldots, N_s^{(i)}-1$, $b=0, 1, 2, \ldots, N_f^{(a)}-1$ and $F_{ab}(i, j)$ is the foreground segmented frame of a shot.

Scenario 1: *To determine left to right movement*

In this scenario, only O_{ab}^{col} is used. If the O_{ab}^{col} throughout a shot is in ascending, then it is asserted that the movement is from left to right. Once the left to right movement is determined, a direction feature vector is loaded with a binary value 1000, i.e. $D^{feat}=\{1000\}$.



7 Spatial feature extraction: (a) keyframe; (b) corresponding re-sampled frame; (c) re-sampled frame alienated into blocks



8 (a) Original frame, (b) foreground extracted frame, (c, d) edge detected frame, (e) triangularised frame and (f) centriod marked frame

Scenario 2: *To determine right to left movement*

Figure 7 describes spatial feature extraction corresponding with keyframe, re-sampled and re-sampled frame alienated in to blocks. Fig. 8 shows the original image in a and in b, c, d, e, f the object is extracted from the frame.

Similar to scenario 1, only $O_{\mathrm{ab}}^{\mathrm{col}}$ is used. If the $O_{\mathrm{ab}}^{\mathrm{col}}$throughout a shot is in descending, then it is asserted that the movement is from right to left. Once movement is determined, the direction feature vector is loaded with 0100, i.e. $D^{\mathrm{feat}}=\{0100\}$.

9 (a) Left to right movement binary value of direction feature vector=1000, (b) right to left movement binary value of direction feature vector=0100, (c) towards the front binary value of direction feature vector=0010 and (d) away from front binary value of direction feature vector=0001

Scenario 3: *To determine towards the front*

In this scenario, $O_{\mathrm{ab}}^{\mathrm{row}}$ is used. If the $O_{\mathrm{ab}}^{\mathrm{row}}$ throughout a shot is in ascending, then it is asserted that the movement is towards front of the screen. Once movement is determined, the direction feature vector is loaded with 0010, i.e. $D^{\mathrm{feat}}=\{0010\}$.

Scenario 4: *To determine away from front*

In this scenario, $O_{\mathrm{ab}}^{\mathrm{row}}$ is used. If the $O_{\mathrm{ab}}^{\mathrm{row}}$ throughout a shot is in descending, then it is asserted that the movement is away from front of the screen. Once movement is determined, the direction feature vector is loaded with 0001, i.e. $D^{\mathrm{feat}}=\{0001\}$.

10 Video frame featuring left to right movement before including the movement feature: (a) frames of the input video sequence; (b) retrieved video frames

11 Video frame featuring right to left movement before including the movement feature: (a) frames of the input video sequence; (b) retrieved video frames

Hence, based on the object movement, features such as centroid feature, distance of movement feature, orientation feature and direction of movement are extracted and stored in the feature library.

Figures 9, 10, 11, 12 and 13 show the movement tracking with this dissimilar frames the movement are identified.

3.4 Occlusion feature extraction

In Ref. 32, the occluded object's boundaries have been detected in the grey-scaled video frames. Converting the video frames from colour to grey scale is a time consuming process. Hence the aforesaid problem is overcome in our proposed system. Occlusion, which occurs between two or more objects, provides insufficient information about object structure and shape, i.e. if the object is retrieved from the occluded objects, then the shape details will be inconsistent.

12 Video frame featuring left to right movement together with the movement feature (our proposed work): (a) frames of the input video sequence; (b) retrieved video frames

13 Video frame featuring right to left movement together with the movement feature (our proposed work): (a) frames of the input video sequence; (b) retrieved video frames

The occlusion feature is detected and recognised by processing the consecutive frames of every shot. The proposed occlusion detection is comprised of three major steps. They are as follows: (1) trigon formation; (2) determining spatial and temporal derivatives; and 3) scum computation.

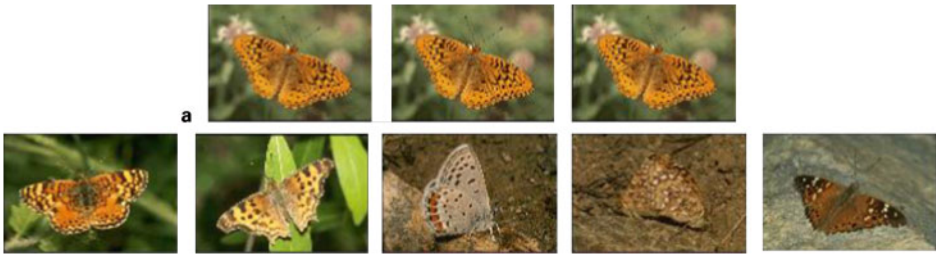
Prior to performing the major steps for detecting the occlusion feature, all the frames are converted from RGB colour space to grey scale frame. The grey scale frames are processed further to determine the occlusion affected area.

3.4.1 Trigon formation

As primary step, a trigon is formed by resizing the frames into different sizes and processed as per the size of the frame. The frame of size $M \times N$ is initially resizing into a frame of size $(M/2) \times (N/2)$. The similar process is repeated by resizing the frame of size $(M/2) \times (N/2)$ into $(M/4) \times (N/4)$. Repeating the process produces the frames with different sizes. When it is arranged in ascending order based on size, it forms a trigon as follows.

To achieve this resizing, a structure-texture decomposition model has been proposed.[29] As per the model, the resizing is performed and finally the trigon of frames is formed.

14 Detection of occlusion in the subjected frames (occluded spot marked with green colour)



15 (a) Frames of input query video and (b) frames retrieved for the object extraction feature without including the spatial feature

3.4.2 *Determining spatial and temporal derivatives*

Spatial derivative is defined as the change of image intensity values as per the change that exhibits in image position. In other words, especially in videos, the spatial derivative is capable of determining the change of intensity values when an object exhibits change in position in two frames. The spatial derivative can be expressed as

$$I'(t,x) = \frac{\partial}{\partial x} I(t,x) \tag{9}$$

The spatial derivative is obtained only in the filtered version of the frame, where the filtering is performed using replicate filtering technique.

The temporal derivative refers to the change of intensity values of an image with respect to a sampling instant of time. It can be defined in other words as the change of pixel intensity values in two different frames. Two different frames refer to the frames that are taken at two different instants of time. The temporal derivative can be given as

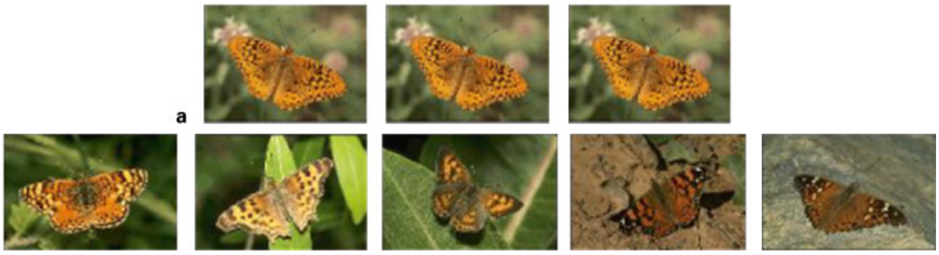
$$I'(t,x) = \frac{\partial}{\partial t} I(t,x) \tag{10}$$

Thus obtained spatial and temporal derivatives from two different frames are utilised further in determining occlusion. The temporal derivates are obtained after applying linear interpolation between two frames.

Figures 15 and 16 show, with the input query video, the frames retrieved for the object extraction feature without including the spatial feature.

3.4.3 *Scum computation*

The indexes of the spatial derivatives and the temporal derivatives are considered in determining the scum of the frames. The scum, here, is represented as the pixel indices which exhibit that motion is determined. The scum is determined by finding the

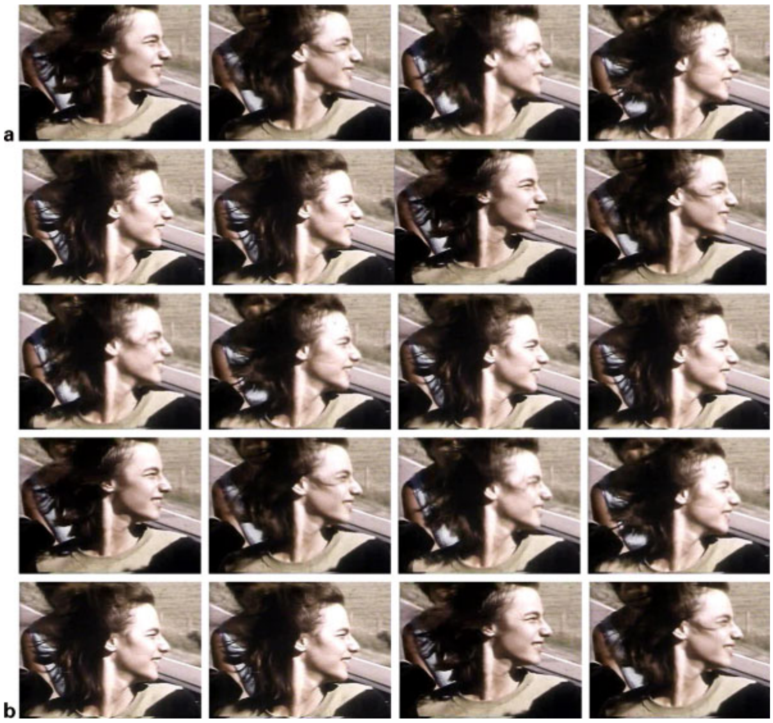


16 (a) Input query video frames and (b) frames retrieved for the object extraction after the inclusion of spatial feature (our proposed work)

17 (a) Frames of the input query video 1 and (b) frames retrieved by our proposed work

pixel indexes which has minimal difference between the pixel locations of the spatial derivative as well as the temporal derivatives. The scum is nothing the occluded portion of the object that is under motion or occluded by the object under motion. This is finally marked and the positions of the occluded pixels and the locations are stored as feature set of the corresponding frame. The feature set is stored in the feature library for the further process of the retrieval stage.

18 (a) Frames of the input query video 2 and (b) frames retrieved by our proposed work



19 (a) Frames of the input query video 3 and (b) frames retrieved by our proposed work

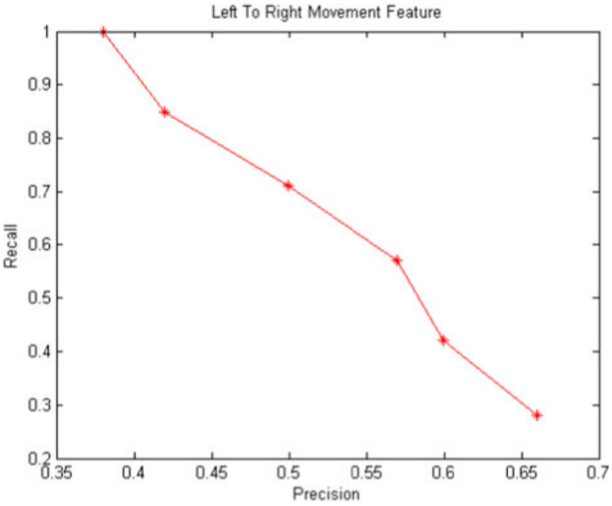
3.5 Retrieval of relevant video clips

The above detailed video clips are subjected to the feature extraction system and so the major feature sets, object-based feature set, movement-based feature sets and occlusion feature sets are extracted. The extracted features are stored in the feature library. When a query clip is given to the CBVR system, the clip is subjected to the feature extraction process and all the aforesaid feature sets are extracted in the similar fashion. Then, each feature set of the database video is analysed by measuring its similarity with the feature set of the database videos. This can be accomplished with the aid of KLD.

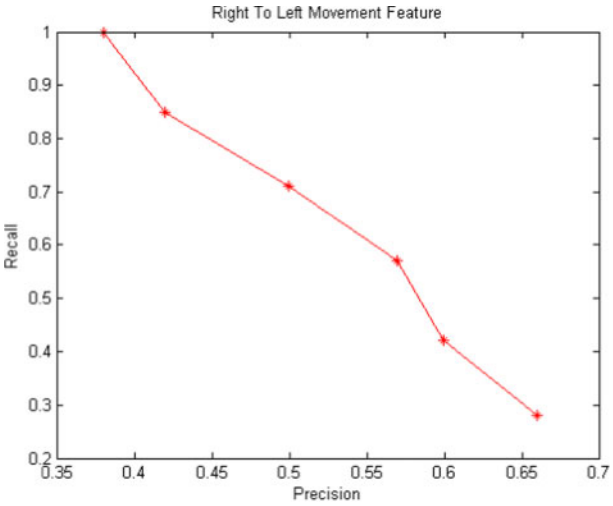
Figures 17–19 show the different input query video frames retrieved by our proposed technique.

KLD-based similarity measure: In Ref. 33, the KLL-based similarity measure has been utilised to compare the sparse multiscale image representations.

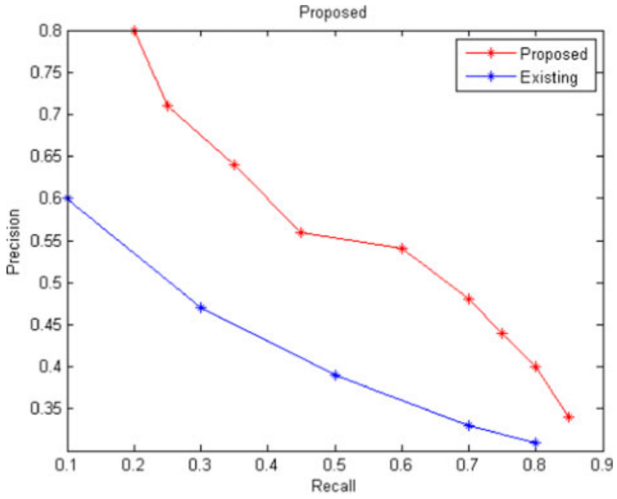
The similarity between the query video clip and each database video clip is determined by calculating



20 Precision-recall values for the left to right movement feature



21 Precision-recall values for the right to left movement feature



22 Comparison of precision-recall values of our proposed work with the existing work[31]

KLD between the feature sets. The KLD can be determined as

$$\mathrm{KLD}(F^{\mathrm{query}},F^{\mathrm{data}})=\sum_{z=0}^{|F^{\mathrm{data}}|-1}F_z^{\mathrm{query}}\log_2\left(\frac{F_z^{\mathrm{query}}}{F_z^{\mathrm{data}}}\right) \quad (11)$$

where $|F^{\mathrm{data}}|$ is the feature set determined for the database video clip. Equation (11) determines the KLD between the feature set of the given query clip and the feature set of a database video clip. In similar fashion, KLD is determined for all the other database video clips. A required number of database video clips, which has minimum KLD, are retrieved as the relevant videos for the given video clip. Hence with the aid of the proposed CBVR system, relevant videos are retrieved from the database effectively.

Figure 22 shows the comparison of precision-recall values of our proposed work with the existing work.

4 IMPLEMENTATION RESULTS AND DISCUSSION

Our proposed CBVR approach has been validated by experiments with a variety of video sequences. The proposed system has been implemented in Matlab (Matlab7.10). We report here some results obtained on a part of a video sequence utilised for retrieval. The results of the object feature, movement feature and occlusion feature are shown below individually.

Compared with existing work,[31] our proposed work has produced better results. Our proposed work has performed satisfactorily when its movement feature was tested on the gait database. After the inclusion of the movement feature, based on the given query video, the frames of the input videos are retrieved.

5 CONCLUSION

A rising research area which has been in limelight recently among the researchers and experimenters is the content-based retrieval of video information. In this paper, an effective content-based retrieval of video has presented which performs proficiently. The first and the foremost process of the proposed technique is that it fragmented the long video sequence into shots. Subsequently, the object, movement and occlusion features are extracted from the obtained shots and these feature sets are kept in the feature library for the subsequent processes. With the aid of the KLD, the similarity measure is evaluated among the features in the feature library and the features of the given query clip are extracted in a similar way. The computed distance decides the retrieval of the similar videos from the collection of videos; hence the effective retrieval of video based on the content has been achieved in an effective manner.

REFERENCES

1 Amiri, A., Fathy, M. and Naseri, A. A novel video retrieval system using GED-based similarity measure. *Int. J. Signal Process. Image Process. Pattern Recognit.*, September 2009, **2**, 99–108.

2 Fan, J. P., Aref, W. G., Elmagarmid, A. K., Hacid, M.-S., Marzouk, M. S. and Zhu, X. Q. MultiView: multilevel video content representation and retrieval. *J. Electron. Imag.*, October 2001, **10**, 895–908.

3 Geetha, P. and Narayanan, V. A survey of content-based video retrieval. *J. Computer Sci.*, 2008, **4**, 474–486.

4 Shanmugam, T. N. and Rajendran, P. An enhanced content-based video retrieval system based on query clip. *Int. J. Res. Rev. Appl. Sci.*, December 2009, **1**, 236–254.

5 Anjulan, A. and Canagarajah, N. Object based video retrieval with local region tracking. *Signal Process. Image Commun.* September 2007, **22**, 607–621.

6 Wu, C.-J., Zeng, H.-C., Huang, S.-H., Lai, S.-H. and Wang, W.-H. Learning-based interactive video retrieval system, Proc. IEEE Int. Conf. on *Multimedia and Expo: ICME '06*, Toronto, Ontario, Canada, July 2006, IEEE, pp. 1785–1788.

7 Sifakis, E., Grinias, I. and Tziritas, G. Video segmentation using fast marching and region growing algorithms. *EURASIP J. Appl. Signal Process.*, 2002, **4**, 379–388.

8 Beevi, Y. and Natarajan, S. An efficient video segmentation algorithm with real time adaptive threshold technique. *Int. J. Signal Process. Image Process. Pattern Recognit.*, December 2009, **2**, 13–28.

9 Rao, D. and Goel, S. Real time retrieval of similar videos in large databases, Proc. Natl Conf. on *VLSI, embedded systems, signal processing and communication technologies: NCVESCOM '09*, Chennai, India, April 2009, AVIT. 1–8.

10 Ye, J., Li, J.-L. and Mak, C. M. Video scenes clustering based on representative shots. *World J. Model. Simul.*, 2005, **1**, 111–116.

11 Tavanapong, W. and Zhou, J. Y. Shot clustering techniques for story browsing. *IEEE Trans. Multimed.*, August 2004, **6**, 517–527.

12 Aslam, N., Irfanullah, Loo, K.-K. and Roohullah. Limitation and challenges: image/video search & retrieval. *Int. J. Digit. Content Technol. Appl.*, March 2009, **3**, 98–102.

13 Dimitrova, N. Multimedia content analysis and indexing for filtering and retrieval applications. *Spec. Issue Multimed. Inform. Technol.*, 1999, **1**, 87–100.

14 Chen, L.-H., Chin, K.-H. and Liao, H.-Y. An integrated approach to video retrieval, *Proc. ACM Int. Conf. Proc. Series*, 2008, **313**, 49–55.

15 Fu, X. and Zeng, J.-X. Local features based image sequence retrieval. *J. Computers*, July 2010, **5**, 987–994.

16 Liu, T. M., Zhang, H.-J. and Qi, F. H. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Trans. Circuits Syst. Video Technol.*, October 2003, **13**, 1006–1013.

17 Avrithis, Y. S., Doulamis, A. D., Doulamis, N. D. and Kollias, S. D. A stochastic framework for optimal key frame extraction from MPEG video databases. *Computer Vis. Image Underst.*, August 1999, **75**, 3–24.

18 Wang, J. Q., Lu, H. Q., Duan, L. Y. and Jin, J. S. Commercial video retrieval with video-based bag of words, Proc. 5th Int. Conf. on *Intelligent multimedia computing and networking: IMMCN '07*, Salt Lake City, UT, USA, World Scientific. July 2007, pp. 1–7.

19 Peng, Y. X., Ngo, C.-W. and Xiao, J. G. OM-based video shot retrieval by one-to-one matching. *Multimed. Tools Appl.*, 2007, **34**, 249–266.

20 Kim, S. H. and Park, R.-H. An efficient algorithm for video sequence matching using the modified hausdorff distance and the directed divergence. *IEEE Trans. Circuits Syst. Video Technol.*, July 2002, **12**, 592–596.

21 Hsieh, J. W., Yu, S.-L. and Chen, Y.-S. Motion-based video retrieval by trajectory matching. *IEEE Trans. Circuits Syst. Video Technol.*, 2006, **16**, 396–409.

22 Zhang, X.-P. and Chen, Z. H. An automated video object extraction system based on spatiotemporal independent component analysis and multiscale segmentation. *EURASIP J. Appl. Signal Process.*, 2006, **2006**, 1–22.

23 Babu, V. and Ramakrishnan, K. R. Compressed domain video retrieval using object and global motion descriptors. *Multimed. Tools Appl.*, January 2007, **32**, 93–113.

24 Wen, C.-Y., Chang, L.-F. and Li, H.-H. Content based video retrieval with motion vectors and the RGB color model. *Forensic Sci. J.*, 2007, **6**, 1–36.

25 Basharat, A., Zhai, Y. and Shah, M. Content based video matching using spatiotemporal volumes. *Computer Vis. Image Underst.*, 2008, **110**, 360–377.

26 Khan, A., Sun, L. F. and Ifeachor, E. Content-based video quality prediction for MPEG4 video streaming over wireless networks. *J. Multimed.* August 2009, **4**, 228–239.

27 Pratheeba, T., Kavitha, V. and RajaRajeswari, S. Morphology based text detection and extraction from complex video scene. *Int. J. Eng. Technol.*, 2010, **2**, 200–206.

28 Hsieh, J.-W., Hsu, Y.-T., Liao, H.-Y. M. and Chen, C.-C. Video-based human movement analysis and its application to surveillance systems. *IEEE Trans. Multimed.*, 2008, **10**, 372–384.

29 Aujol, J.-F., Gilboa, G., Chan, T. and Osher, S. Structure-texture image decomposition – modeling, algorithms, and parameter selection. *Int. J. Computer Vis.*, April 2006, **67**, 111–136.

30 Rajendran, P. and Shanmugam, T. N. A content-based video retrieval system: video retrieval with extensive features. *Int. J. Multimed. Intell. Secur.*, 2011, **2**, 146–171.

31 Hu, R. and Collomosse, J. Motion-sketch based video retrieval using a Trellis Levenshtein distance, Proc. Int. Conf. on *Pattern recognition: ICPR '10*, Istanbul, Turkey, August 2010, IEEE Computer Society, pp. 121–124.

32 Stein, A. N. and Hebert, M. Local detection of occlusion boundaries in video, Proc. BMVC 2006, Edinburgh, UK, September 2006, BMVA, pp. 407–416.

33 Piro, P., Anthoine, S., Debreuve, E. and Barlaud, M. Image retrieval via Kullback-Leibler divergence of patches of multiscale coefficients in the KNN framework, Proc. Int. *Workshop on Content based multimedia indexing: CBMI '08*, London, UK, June 2008, IEEE, pp. 230–235.

34 Padmakala, S., AnandhaMala, G. S., and Shalini, M. An effective content based video retrieval utilizing texture, color and optimal key frame features, Proc. Int. Conf. on *Image information processing: ICIIP '11*, Himachal Pradesh, India, November 2011, IEEE, pp. 1–6.

35 Gupta, S., Gupta, N. and Kumar, S. Evaluation of object based video retrieval using SIFT. *Int. J. Soft Comput. Eng. (IJSCE)*, May 2011, **1**, 1–6.