Video Indexing and Retrieval

Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 2.6.1

Artificial Intelligence & Information Analysis Lab

Video Indexing and Retrieval

- Hierarchical video structure
- Shot cut/transition detection
- Video Summarization
- Audiovisual content description
- Video indexing and retrieval

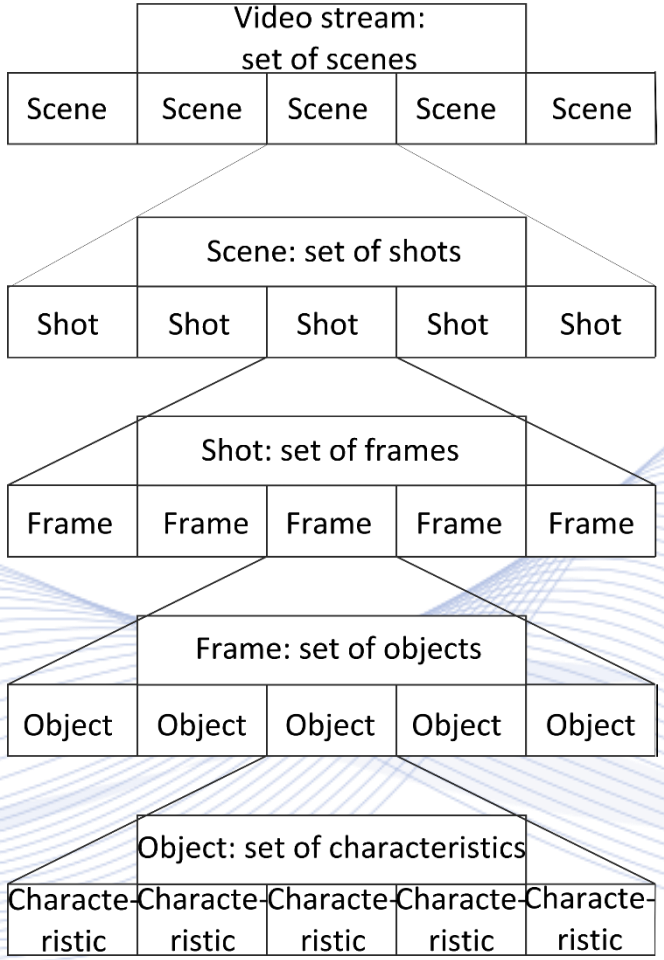
Introduction

- Video content search:
  - in a digital video file;
  - in broadcasting archives;
  - in social media sites (e.g., YouTube).
- Content-based video search and retrieval is difficult:
  - Too big content size;
  - Unstructured video content
  - Time consuming video browsing.

Introduction

- Many techniques have been proposed for content-based video indexing and retrieval:
  - Shot cut/transition detection
  - Video Summarization
  - Video key-frame selection
  - Audiovisual content description
  - Video indexing and retrieval.
- Low-level and semantic (content-based) video retrieval techniques.

Hierarchical video structure



Hierarchical video segmentation.

Hierarchical video structure

- A video (e.g., a movie) consists of a sequence of scenes.
- A **video scene** is a sequence of video shots focusing on an object or objects or story of interest.
- A **video shot** is a single sequence of frames which are captured by a stationary or continuously moving camera.
  - A movie which contains alternating views of two persons consists of multiple shots.

Hierarchical video structure

Example:

- a person walking in a corridor and entering a room could constitute a scene, if it has been captured by multiple cameras.
- Three camera shots, showing three different persons walking down a corridor may constitute a scene, if the object of interest is the corridor and not the persons themselves.

Shot cut and transition detection

There are various types of **shot transitions**:

Abrupt shot transitions:

- A **shot cut** is an abrupt shot change.
- Abrupt changes are easier to detect, compared to gradual ones.

Gradual shot transitions:

- A **fade-in/fade-out** is a slow change in shot luminance, which usually leads to, or starts with, a black frame.

Shot cut and transition detection

- A **dissolve** takes place when there is a spatial overlay of the frames of the two shots for the duration of the transition. The luminance of the images of the first shot decreases and that of the second shot increases.
- A **wipe** occurs when the pixels of the second shot gradually replace those of the first shot with a local motion, e.g., from left to right.

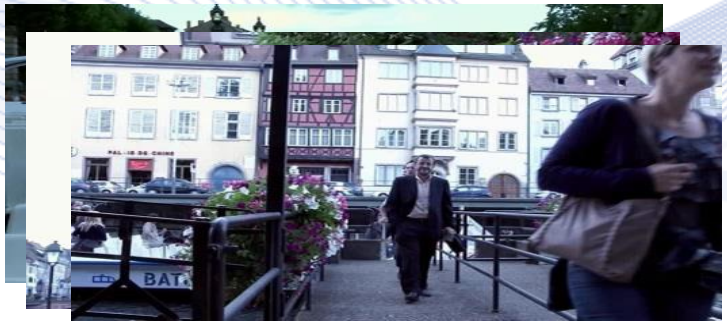
- Many other gradual shot transition types are possible.

Shot cut and transition detection

Temporal Decomposition of a video into shots.

movie

shots



Shot cut detection

Let video frames f, f' have luminance vectors $\mathbf{Y}, \mathbf{Y}'$.

- The simplest **distance metric** between two consecutive shots is given by:

$$D(f, f') = \|\mathbf{Y} - \mathbf{Y}'\|.$$

- Various error norms, e.g., the L_1 or L_2 ones can be used.
- This shot cut detection method has a limited success rate and can detect only 73% of the actual shot changes.

- False detections occur in case of object or camera motion.

Shot cut detection

Histogram-based shot cut detection:

- **Image histogram** $\hat{p}_f$ *is estimated by:*

- $\hat{p}_f(f_k) = \frac{n_k}{n}, \quad k = 0,1,\ldots,L-1.$
- n: total number of image pixels
- n_k, $k = 0,1,\ldots,L-1$: number of image pixels having intensity.

- The simplest, most efficient and most commonly used shot cut detection methods are based on histograms and their variants.

Shot cut detection

Histogram-based shot cut detection:

- The histogram of a video frame is calculated and compared to that of the next frame.
- The histograms of two video frames belonging to different shots are different.
- The histograms of the frames in a shot must be similar to each other in the case of camera motion or object motion:
  - the histogram is not sensitive to the object position inside a video frame.

Shot cut detection

- We can use the one-dimensional luminance histograms $\boldsymbol{h}_f$ or two-dimensional chrominance histograms, or three-dimensional histograms of a video frame f in a desired color coordinate system.

- Alternatively, three one-dimensional histograms can be used, one for each color coordinate, e.g., R, G, B.

Shot cut detection

- Let $h_f(k)$, $0 \leq k \leq 255$ be the values of the luminance histogram for frame f (for 8 bit luminance). Several variants of the histogram method are described subsequently.

- It is remarkable that, depending on the application, these methods were able to detect over 95% of shot changes in various video data.

Shot cut detection

- The simple histogram difference between the frames f and $f'$is given by :

$$D(f, f') = \sum_{k=0}^{K} \left| h_f(k) - h_{f'}(k) \right|.$$

- A larger weight can be assigned to a dominant color in the comparison between two frame histograms. This is done by using the weighted histogram difference defined as :

$$D(f, f') = \frac{R}{S} * D_R(f, f') + \frac{G}{S} * D_G(f, f') + \frac{B}{S} * D_B(f, f'),$$

$$S = (R + G + B)/3.$$

where R, G and B are the red, green and blue channel intensities and D_R, D_G, D_B are the corresponding histogram differences.

Shot cut detection

- The equalized histogram difference uses equalized frame histograms :

$$D(f, f') = \sum_{k=0}^{K} \left| h_{ef}(k) - h_{ef'}(k) \right|.$$

- The aim of histogram equalization is to produce a uniformly distributed histogram $\boldsymbol{h}_{ef}$ for each video frame.

Shot cut detection

- Histogram intersection can be used for measuring video frame similarity :

$$D(f, f') = \sum_{k=0}^{K} \min\left(h_f(k), h_{f'}(k)\right).$$

- The intersection of two similar histograms is maximal (equal to 1.0), while, in dissimilar frames, this value is generally much smaller.

Shot cut detection

- The squared histogram difference attempts to smooth large histogram differences in two frames :

$$D(f, f') = \sum_{k=0}^{K} \frac{(h_f(k) - h_{f'}(k))^2}{h_f(k)}.$$

- Division by $h_f(k)$ is used as a normalization factor. A more efficient variant is the following one :

$$D(f, f') = \sum_{k=0}^{K} \frac{(h_f(k) - h_{f'}(k))^2}{max(h_f(k), h_{f'}(k))}.$$

Shot cut detection

- Shot cut methods based on the rate of pixel value changes, model the difference between the pixel values in two consecutive frames as a combination of three factors :

  a) Pixel value change, due to object or camera motion, or focus and illumination change at a given instance in a given video frame.

  b) Change resulting from a cut, wipe, dissolve or fade-out.

  c) A small additive zero-mean Gaussian noise, which models the camera and digitization noise.

Shot cut detection

- This model leads to different luminance differences between two video frames for cuts, dissolves and wipes, respectively.

- Shot cut detection based on edge detection [Hanjalic99] uses the observation that, during shot change, new image edges appear away from the positions of old ones.

- Therefore, shot cut can be identified by comparing the edges of two consecutive video frames.

Shot cut detection

- If p_1, p_2 denotes the percentage of edge pixels (in frame f, f'), which are spatially further apart than a predefined distance d from the closest edge pixel in the frame f', f respectively, the dissimilarity between two frames f, f' is given by :

$$D(f, f') = \max(p_1, p_2).$$

Shot transition detection

- A method which has been proposed for both abrupt and gradual shot transition detection uses the joint entropy between two video frames f, f' [CER06] :

$$H(f,f') = -\sum_{f\in f', f'\in f'} P_{FF'}(f,f') log P_{FF'}(f,f'), \quad (15.3.9)$$

- or the respective mutual *information* :

$$H(f,f') = -\sum_{f\in f', f'\in f'} P_{FF'}(f,f') log \frac{P_{FF'}(f,f')}{P_F(f)P_{F'}(f')}. \quad (15.3.10)$$

Shot transition detection

- To find their luminance similarity. Here F, F' are considered to be random variables, which correspond to the luminances of the two video frames $\boldsymbol{f}$ and $\boldsymbol{f'}$. $P_F(f), P_{F'}(f'), P_{F,F'}(f, f')$ are the marginal and joint luminance distributions of the two frames.

Shot transition detection

- The ratio of incremental gradual changes in the mean signal luminance to the corresponding changes of the chrominance signal is considered as the criterion for the fade-out detection. [Fernando99]

- The mean chrominance value varies less than the mean luminance value for a video sequence without a fade-out. During shot fade-out, both mean values vary equally. By selecting a predefined threshold for their ratio, the shot fade-out can be detected.

Shot transition detection

- Another approach, which detects all gradual shot transitions, is based on modeling the transition pattern [Bescos00]. The difference between the histogram of the i-th frame being processed and the histogram of the $(i - l)$-th frame is calculated.

- By setting I equal to the length of the gradual transition L (expressed in frame number), the shot transition can be found. These parameters (I and L) can be selected according to the various gradual shot transition types.

Shot transition detection

- A method based on the *fusion* of the results of more than one shot transition detection techniques was proposed in [Yusoff99]. The overall shot transition detection performance can be improved, and the error rate can be reduced by 50% by fusing the individual detection results.

Key frame selection and video summarization

- Consider we want to extract a number of *key frames*, able to summarize well the video content, for video description and fast browsing [Shan98], in a long video sequence. Their number may vary from 5% to 10% of the total frame number in the original video.

- There is no mathematical model, which defines the exact requirements for key frame selection. Many techniques are based on shot cut detection, while other approaches employ the visual content and motion analysis.

Key frame selection and video summarization

- A technique which involves the use of color histograms in the *YUV* space has been proposed in [Gunsel98] and has been improved in [GQI00]. Let $D(f_t, f_{t+1})$be the difference of the one-dimensional YUV histograms :

$$D(f_t, f_{t+1}) = \sum_{k=0}^{K} |h_{t+1}^{Y}(k) - h_t^{Y}(k)| + \sum_{k=0}^{K} |h_{t+1}^{U}(k) - h_t^{U}(k)| + \sum_{k=0}^{K} |h_{t+1}^{V}(k) - h_t^{V}(k)|.$$

where $h_t^Y(k)$, $h_t^U(k)$, $h_t^V(k)$ denote the histogram of the Y, U, V color coordinates of frame t.

Key frame selection and video summarization

- Let $D(f_t, \bar{f}_n)$ denote the difference between the color histogram of the current frame t and the mean histogram of the preceding n frames :

$$D(f_t, \bar{f}_n) = \sum_{k=0}^{K} |h_t^Y(k) - \bar{h}_n^Y(k)| + \sum_{k=0}^{K} |h_t^U(k) - \bar{h}_n^U(k)| + \sum_{k=0}^{K} |h_t^V(k) - \bar{h}_n^V(k)|$$

where $\bar{h}_n^Y$, $\bar{h}_n^U$, $\bar{h}_n^V$ are the mean histograms of the preceding n frames $f_{t-1}, \dots, f_{t-n}$:

$$\bar{h}_n^Y(k) = \frac{1}{n}\sum_{j=1}^{n} h_{t-j}^Y, \qquad \bar{h}_n^U(k) = \frac{1}{n}\sum_{j=1}^{n} h_{t-j}^U, \qquad \bar{h}_n^V(k) = \frac{1}{n}\sum_{j=1}^{n} h_{t-j}^V$$

Key frame selection and video summarization

- A frame becomes a key frame if $D(f_t, f_{t+1}) > T$, where T is the scene change detection threshold.
- If $D(f_t, f_{t+1}) < T$, then $D(f_t, \bar{f}_n)$ is checked.
- If $D(f_t, \bar{f}_n) > T$, then the frame also becomes a key frame.
- The smaller T is, the higher is the number of the selected key frames. An automatic threshold selection is presented in [GQI00].

Key frame selection and video summarization

- Another approach, is based on low motion and high spatial activity and the presence of human faces [Dirfaux00]. A score $S_c(f)$ is calculated for frame f, using the following equation:

$$S_c(f) = W_H(H(f)/\sigma_H) + W_S(S(f)/\sigma_S) + W_F(F(f)/\sigma_F) - W_D(D(f)/\sigma_D).$$

$H(f)$ is the entropy of the frame f:

$$H(f) = -\sum_{f \in \boldsymbol{f}} p_F(f) \log p_F(f),$$

Key frame selection and video summarization

- $p_F(f)$ is the probability density function of the frame luminance f.
- $S(f)$ and $F(f)$ are the skin color and face detection functions, respectively.
- $D(f)$ is the pixel-based frame difference.

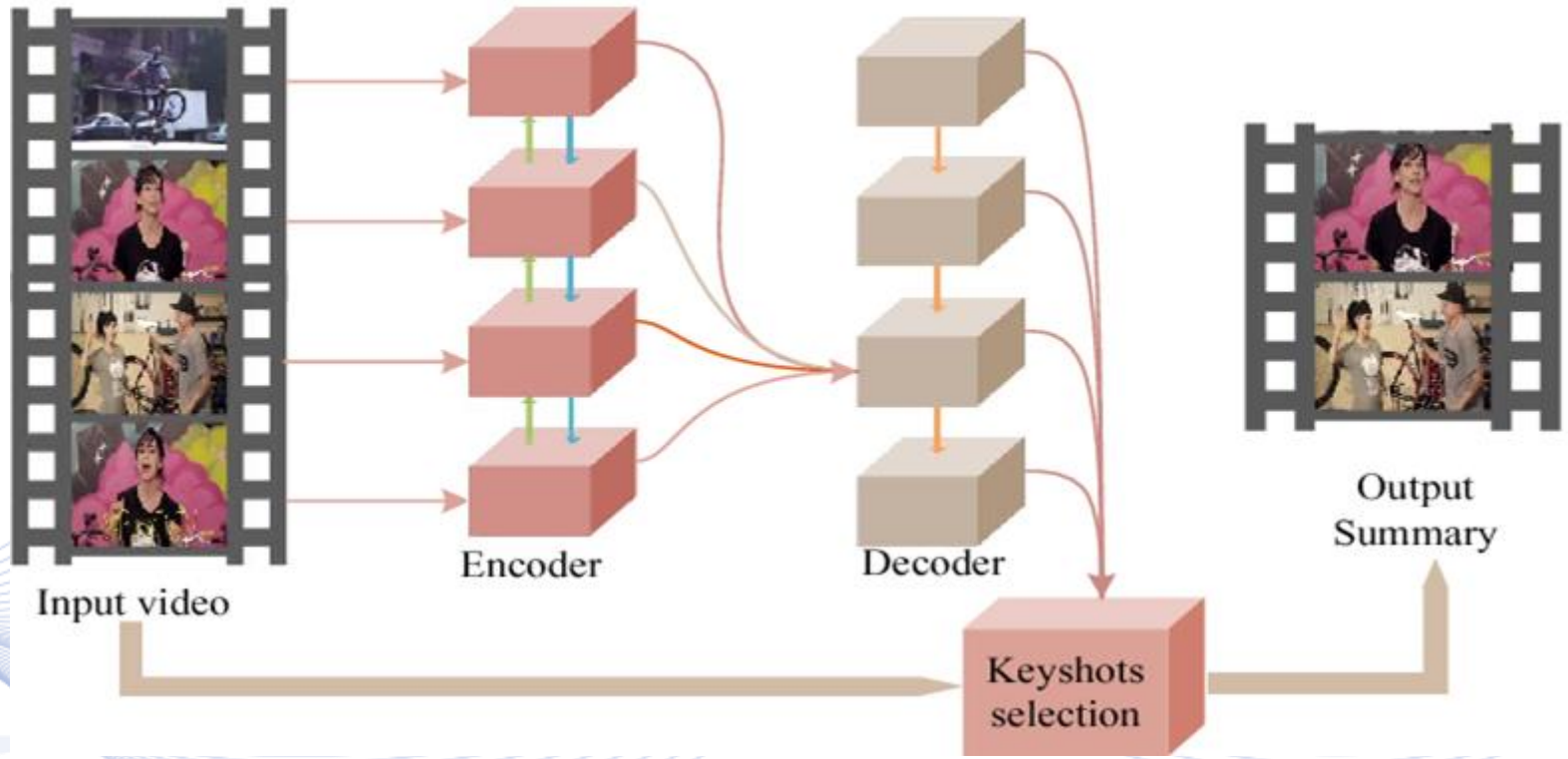
Key frame selection and video summarization

- W_H, W_S, W_F, W_D are heuristically found weighting factors

- σ_{H}, σ_{S}, σ_{F}, σ_{D} are the standard deviations of frame entropy, skin color, face and frame difference, respectively.

The frame with the highest score is selected as a key frame. This method can be generalized for the selection of many key frames.

Key frame selection and video summarization



(Picture from CatalyzeX)

Object based shot description

- Many techniques have been proposed in the literature for video summarization and retrieval based on object detection.
- The objects can be detected using spatial features, such as their color, texture or shape.
- In a video, there are two sources of information which can be used for object detection and tracing : visual characteristics and motion information.

Object based shot description

- A typical strategy is to initially perform region segmentation based on color, texture and shape information.

- After the initial segmentation, regions with similar motion vectors can be merged based on certain limitations, such as region adjacency [Aslandogan99].

Object based shot description

A method which incorporates color, shape and spatial analysis for video summarization was proposed in [HUNG00]:

- For each image (or key-frame), the color histogram is produced in the RGB space.
- The shape information is determined by the edge direction histogram.
- Edges are produced by the Sobel edge detector [PIT01].

Object based shot description

- To differentiate between two images with similar color and shape information, the image is partitioned in nine rectangular regions.

- The choice of the number of regions had been made experimentally. Color and shape analysis is performed in each region.

Object based shot description

- To retrieve stored images, the input image is processed in the same way as the stored images.
- The absolute difference between the descriptions of the input image and the stored images and the images with the greater similarity are selected.

Object based shot description

A pattern recognition method employing the class centroids is used for video segmentation based on object similarities [KIM01].

- The similarity measure used is the Euclidean distance between two object masks.
- The frames which contain similar objects are grouped as a class or as a shot.
- The median frame of the class is considered the key-frame.

Object based shot description

Another method for object detection in video uses subsets of pixels which do not follow the total motion [Coudert99].

- A frame is transformed into two one-dimensional discrete signals, using a Mojette transform [Zhuang98].

Object based shot description

- The motion estimation procedure is utilized and a cross-correlation coefficient results.

- A low cross-correlation coefficient from each 1-D signal is selected and is backprojected on the original 2-D frame for the production of a mask, which represents the object (or region) of interest.

Object based shot description

A method for the segmentation of moving scenes using a Median Radial Basis Function (MRBF) neural network has been proposed in [Bors97].

- The image is divided in blocks on a rectangular grid. A five-dimensional feature vector is assigned to each block.

Object based shot description

- This feature vector contains the (x, y) position coordinates (two features), the luminance and local motion vector coordinates (two features)

- By estimating the Radial Basis Function [Bors96], each block is assigned to a moving object region, which corresponds to the most active RBF. When training finishes, the maximally activated outputs indicate the corresponding moving objects.

Object based shot description

The use of the adaptive K-means (C -means) algorithm is used in [Kompatsiaris01] for connected region (object) detection.

- A specific number of consecutive video frames are processed in the *CIE* color space.
- The technique starts with the segmentation of each frame in K subdivisions, using color histograms.

Object based shot description

- It iteratively segments the regions using color difference, motion and spatial distance parameters, until a connectivity limitation parameter is satisfied and results to an object.

- If the connectivity limitation is not satisfied, the technique shifts to another set of predefined chromatic difference, motion and spatial difference parameters.

Object based shot description

Genetic Algorithms *(GA)* were presented in [Hwang01] for object detection in video.

- The recommended procedure consists of three steps: spatial segmentation, motion estimation and video object detection.
- This approach is more suitable for off-line processing, due to its high computational complexity.

Multimodal audiovisual content description

The combination of audio and image features to extract more semantic features was proposed in [Adami01]:

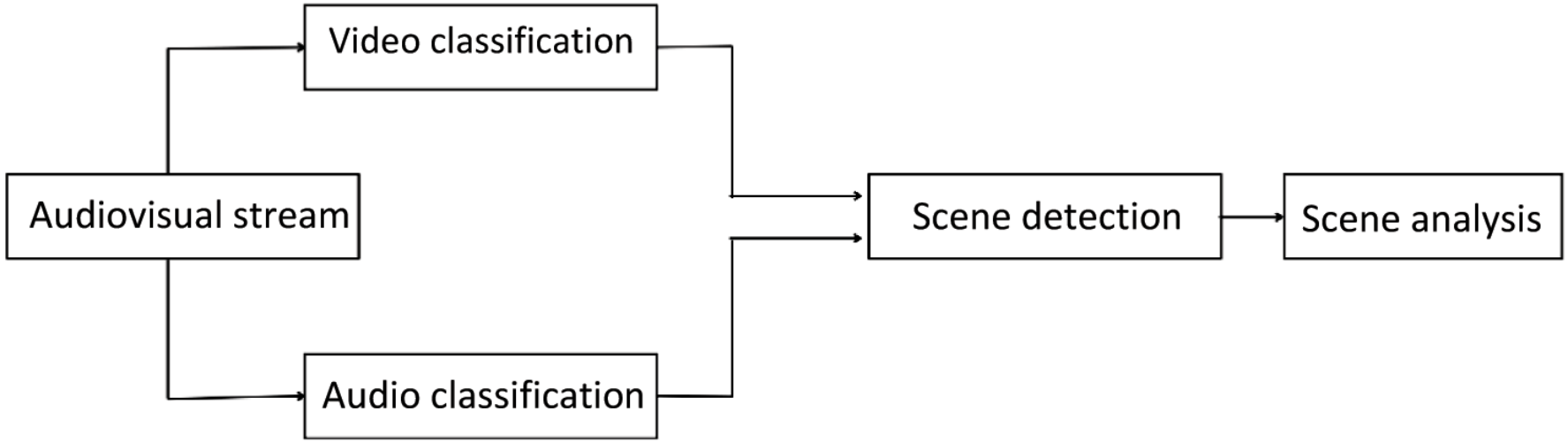
- The mean audio intensity is used as a measure of shot significance
- For audio streams, the audio features are extracted from low-level audio properties.
- For video streams, the visual features are extracted using motion estimation with luminance histograms and pixel differences.

Multimodal audiovisual content description

- Each sequence of features extracted from both audio and video channels is used for the identification of video semantics.

- In case of scenes containing humans, four different shot types can be identified: dialogue, monologue, action and generic video.

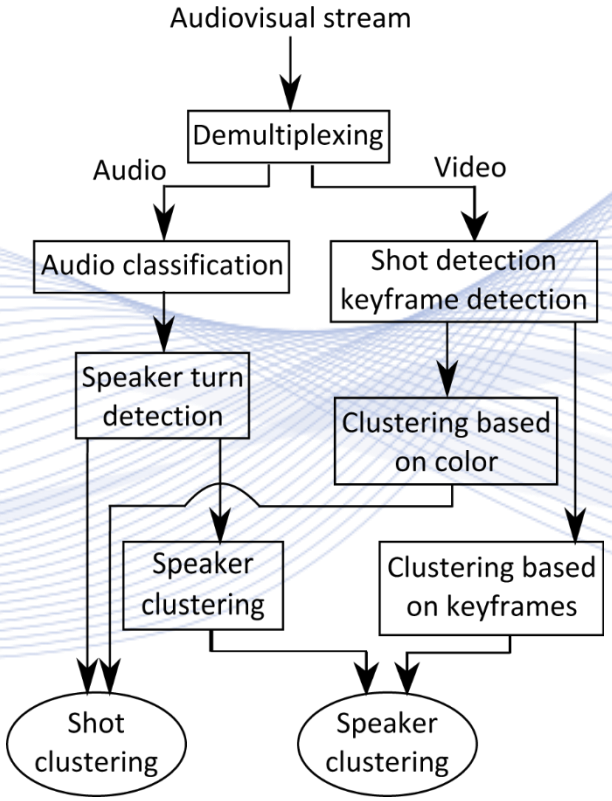
Multimodal audiovisual content description



Combination of audio and video features for the description of audiovisual content.

Multimodal audiovisual content description

- News video description techniques, using audio have been proposed in [WQI00] and have been further improved with the addition of text processing in [Jiang00]. Figure 3 shows the diagram of the proposed method.



Audio assisted news video processing method.

Multimodal audiovisual content description

- The news video stream input is split in an audio and a video stream.
- Subsequently, the audio stream is classified in four classes: speech, music, environment noise and silence.
- The audio stream is further segmented in different segments, depending on the current speaker. Simultaneously, video shot detection and key-frame extraction are performed in the video stream.

Multimodal audiovisual content description

- The color correlation between shots is calculated and an extended window grouping algorithm is performed, so that shots, whose objects or the background are closely correlated are grouped together.

- In the next step, the results of audio and video analysis are merged to improve the shot grouping.

- At this point, the results of speaker change detection are combined with the results of color-based shot grouping, to find the scene transitions

Multimodal audiovisual content description

- The fusion rule is that shots in a speaker segment, which are correlated according to color analysis, are grouped and marked as correlated.

- In other words, a shot sequence will be grouped in a scene only when the correlation analysis of the visual content and the audio segmentation detect a common scene transition.

Multimodal audiovisual content description

For transition detection in news videos, a robust detection of anchorpersons is required.

- As far as audio is concerned, the audio segments are marked by the speaker change points and are further grouped.

- As far as video is concerned, the shots are grouped based on key-frame clusters.

Multimodal audiovisual content description

The following heuristic rules are used for the detection of potential key persons in audio and video data :

a) The speech/image duration ratio of anchorpersons in news broadcasts is usually higher than the corresponding speech/image duration ratio of other persons.

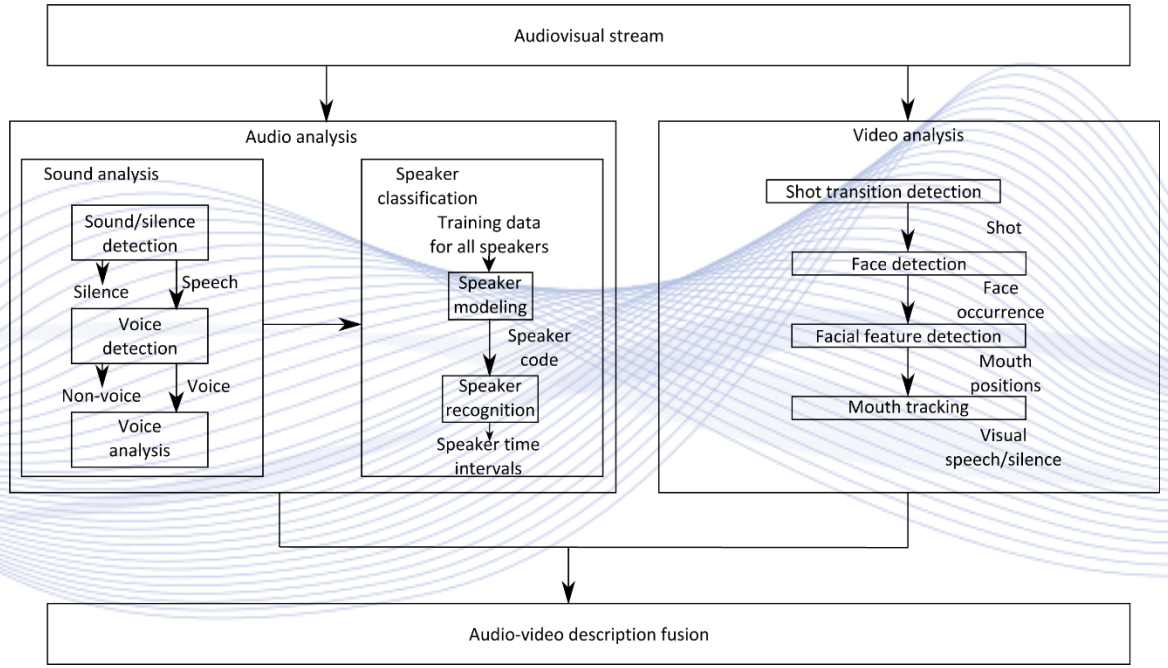
b) The temporal dispersion of the speech/image duration ratio of the anchorpersons is higher than the corresponding speech/image duration ratio of other persons. That is, the anchorperson will appear from the start until the end of a news broadcast.

Multimodal audiovisual content description

In this case, the visual and audio analysis results are recombined, resulting to a valid anchorperson shot detection. The fusion rule used here is the logical conjunction between their video shot durations and their speech segments.

Multimodal audiovisual content description

- A method involving both visual and audio content for video indexing was presented in [TSE99], [TSE01]. The block diagram of this method is shown in Figure 4.



Audiovisual content analysis.

Multimodal audiovisual content description

- Audio processing starts with speech-silence detection which is performed using the mean volume of the audio signal and the zero crossings rate, leading to signs of potential scene changes.

- The audio features are extracted only from voiced frames, therefore, the video frames must be divided in voiced and unvoiced ones.

- This is done using the low to high frequency ratio of the Short Term Fourier Transform *(STFT)* of the audio channel. The extracted features can model well and identify the speaker.

Multimodal audiovisual content description

- Video processing starts with the detection of shot transitions, using color differences. Subsequently, face detection is performed, assuming that faces can be characterized by skin-like color and have elliptical shape.

- Facial analysis follows, to estimate the mouth position. Mouth tracking is used to determine if a person speaks or not. Cross-modal audio and video analysis facilitates the detection of shot transitions, using the detection of speech-silence periods. Additionally, face analysis facilitates speaker identification.

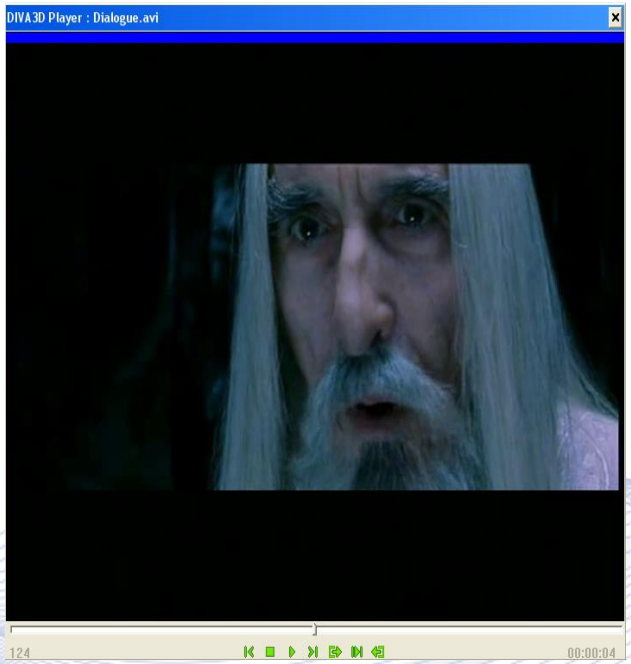
Semiautomatic video description and search approaches

- Semiautomatic approaches for content-based video retrieval were proposed in [LIU00] and [OH00].

Most videos are annotated manually. Different people may have different semantic interpretations for the same video scene.

- To eliminate this confusion, an adaptive and flexible automatic approach was proposed [LIU00].

Multimodal audiovisual content description



Detection of: a) dialogs in a movie and b) monologues in news broadcasting.

Semiautomatic video description and search approaches

- Each time a user searches for a video with text-based query, text annotation tables are searched.
- If the text in the query is not found, the video clip is reprocessed in real time for possible extraction of the query text.
- When the query is satisfied, the query terms are added to the annotation text.
- Using this combination of retrieval and processing, the system can include different descriptions for the same video clip, preserving at the same time semantic soundness.

Semiautomatic video description and search approaches

- Another approach, which utilizes human interaction to simplify retrieval and improve performance, preserving at the same time content integrity was presented in [OH00].

- Human interaction is needed for the selection of a single scene.

- The technique shall select other key scenes automatically to compile a complete video clip summary. It requires limited time and effort and results in higher performance in the compilation of video summaries.

Indexing Techniques

- In the previous sections, techniques were presented for analyzing video streams and extracting semantic information.

- The next step is to present techniques for creating the appropriate indexing structure of the acquired information to facilitate video retrieval. *Hashing* is a widely known technique for data indexing [KNU73].

- An approach for video image (key-frame) indexing based on the edge directions in predefined image regions has been presented in [MOT00].

Indexing Techniques

- The image is divided in 8×8 regions. For each region, the histograms of four edge directions (horizontal, vertical and two diagonal) are calculated using the edge operator discussed in [CAN86], which results in a four bin histogram.

- The histogram is normalized by the image size to operate on images of various sizes. Subsequently, the value of each histogram bin is quantized in two quantization modes: one with five values and one with seven values.

Indexing Techniques

- This quantization serves two purposes:

a) it reduces the number of records in the hash table and

b) assigns the same quantized value to similar images.

- To circumvent the problem of having two similar values in the histogram bin (on just below the quantization level and one just above) to correspond to different quantization levels, the bin values are quantized twice using overlapping quantization intervals.

Indexing Techniques

- The address in the hash table is created, as shown in Figure 5. There are 64 regions in each image, requiring 6 bits for their representation.

- For each image region, there are two modes of quantization for each one of the four histogram bins. A 4-bit flag indicates which mode is used for each of the four bins (one flag for each bin).

- Finally, four 3-bit quantization values for each one of the four edge directions (histogram bins) are used.

Indexing Techniques

Block	Bin quantization	Quantization level for each bin
6 bits	4 bits	12 bits

Address generation in the hash table.

Indexing Techniques

- By this way of histogram quantization, we expect that visually similar images correspond to the same hash table records.
- Each record may have multiple indices to the images which are stored in the database.
- When the database is queried using an image as the query example, the query image is subjected to analysis, using the same procedure and images which have the same index in the hash table are retrieved.

Indexing Techniques

- Geometric hashing has attracted many researchers working in content based indexing [WANG99,MAH00,GAV92,LAN88].

- Geometric hashing has many advantages, thus rendering it suitable for video indexing. It is not sensitive to object rotations and translations. Additionally it is very efficient, has a low polynomial complexity and is inherently parallel. The general approach for the creation of hash tables works as follows.

Indexing Techniques

- For each image, a feature set is determined. Subsequently, these features are extracted and mapped to two-dimensional position coordinates.
- The appropriate feature mapping model depends on the application at hand. An initial position point is selected and the rest of the points are positioned sequentially, in a clockwise manner.
- For each ordered feature pair, their distance is calculated and quantized. A hash table record is constructed using all the quantized distance values.
- A description of geometric hashing can be found in [WOL97].

Indexing Techniques

- A three-dimensional method for extendible hashing uses multi-precision color similarity in the RGB color space [LIN01].

- It is applicable on other color spaces as well. It is an extension of linear hashing, that avoids retaining much hash table information, providing at the same time the ability for multi-precision color matching.

- The hash table can be extended along the three RGB dimensions.

Indexing Techniques

- A tracking mask is required to track the split address and check the hash table growth by mapping three dimensions to one.

- The hash table has three initial depths (d_1, d_2, d_3), one for each R, G, B color coordinate.

- It also has a growth depth d_g, which is initially zero and increases in the address space.

Indexing Techniques

- The number of bits of a hashing address is $(d_1+d_2+d_3+d_g)$. This way, the hash table has $2^{(d_1+d_2+d_3+d_g)}$ records.

- A segment is a group of records which has the same p_1, p_2, p_3 most significant bits of the RGB values, respectively. These bits (p_1, p_2, p_3) are called local depths. Initially, the local depths of the segments are the same as the initial list depths.

Indexing Techniques

- The hash table record either points to a segment or has a zero value, if there are no records there. The initial hash table has $2^{(d_1+d_2+d_3+d_g)}$ records, as the depth d_g is zero initially.

- When a segment overflows and has to be split, the local depths are compared to the table depths. If $p_1 + p_2 + p_3 = \mathrm{d}_1 + d_2 + d_3 + d_g$, then there is no address space to store the new segment.

- The hash table must be doubled and d_g is increased by 1. Otherwise, the address space remains unchanged. The initial segment has the same hash address k, as before.

Indexing Techniques

- The hash address of the new segment is $2^{p_1+p_2+p_3} + k$. The segment is divided along the dimension of the highest variance.

- Let us assume that the R dimension has the highest variance. The segment is split, moving all records, whose the $(p_1 + 1)$ initial bit of the R value is 1, to the new segment. The local depths of the two split segments are now $(p_1 + 1, p_2, p_3)$.

- The k-th record of the $(p_1 + p_2 + p_3 - d_1 - d_2 - d_3 + 1)$-th level of the tracking mask is renewed, to point to the split.

Indexing Techniques

- In this hash table structure, a given color can be represented as follows :

1. Extraction of the d_1, d_2, d_3 most significant bits from the R, G, B pixel values, respectively. Use of these bits for the creation of an original hash table address.

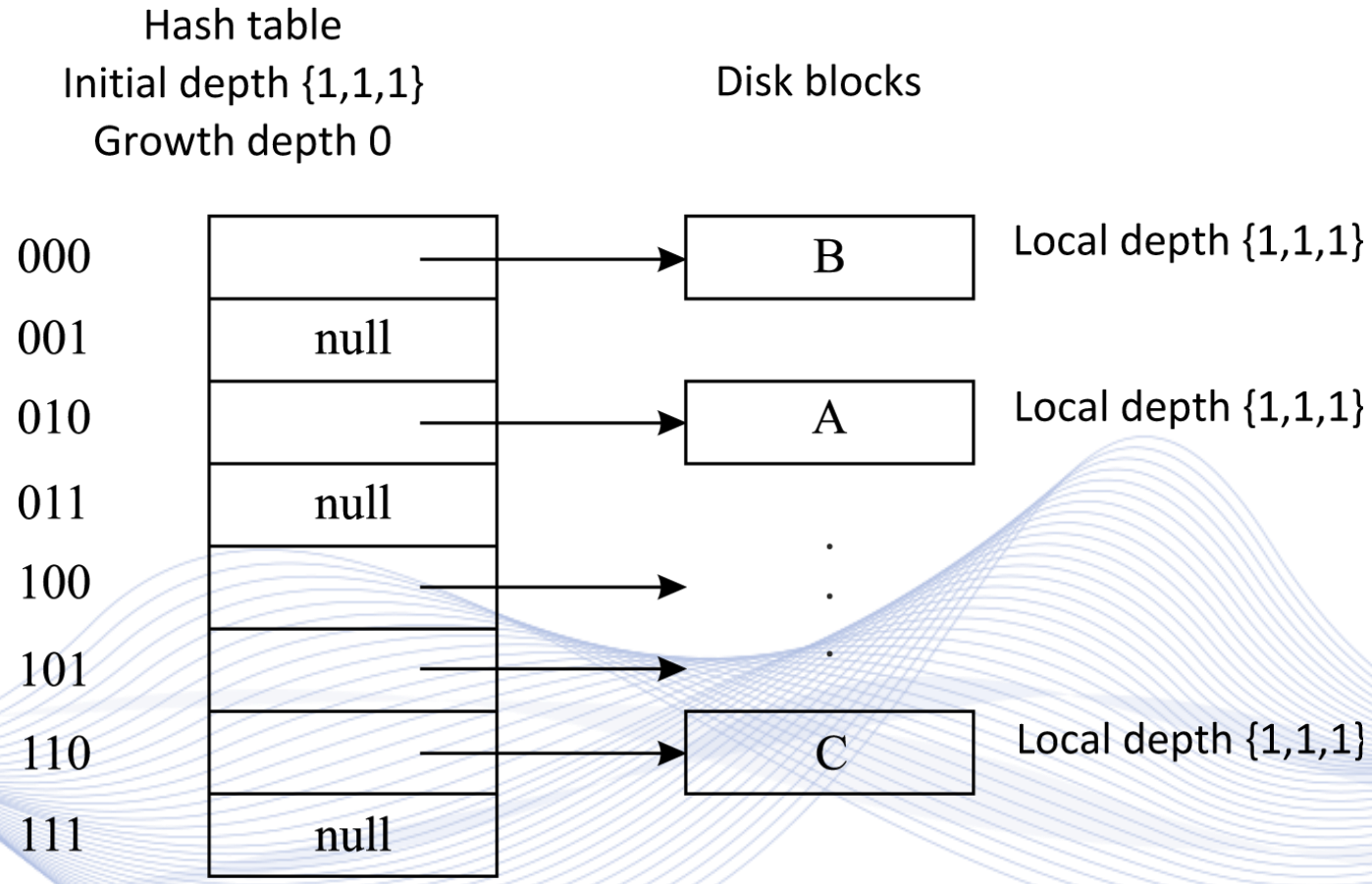
2. Reference to the record which corresponds to the initial address in the first level of the tracking mask. If the segment has not been split, go to step 4, otherwise calculate the new address, as described above.

Indexing Techniques

3. Tracking of the next level of the tracking mask with the new address, until the mask indicates that the segment represented by the address has not been split.

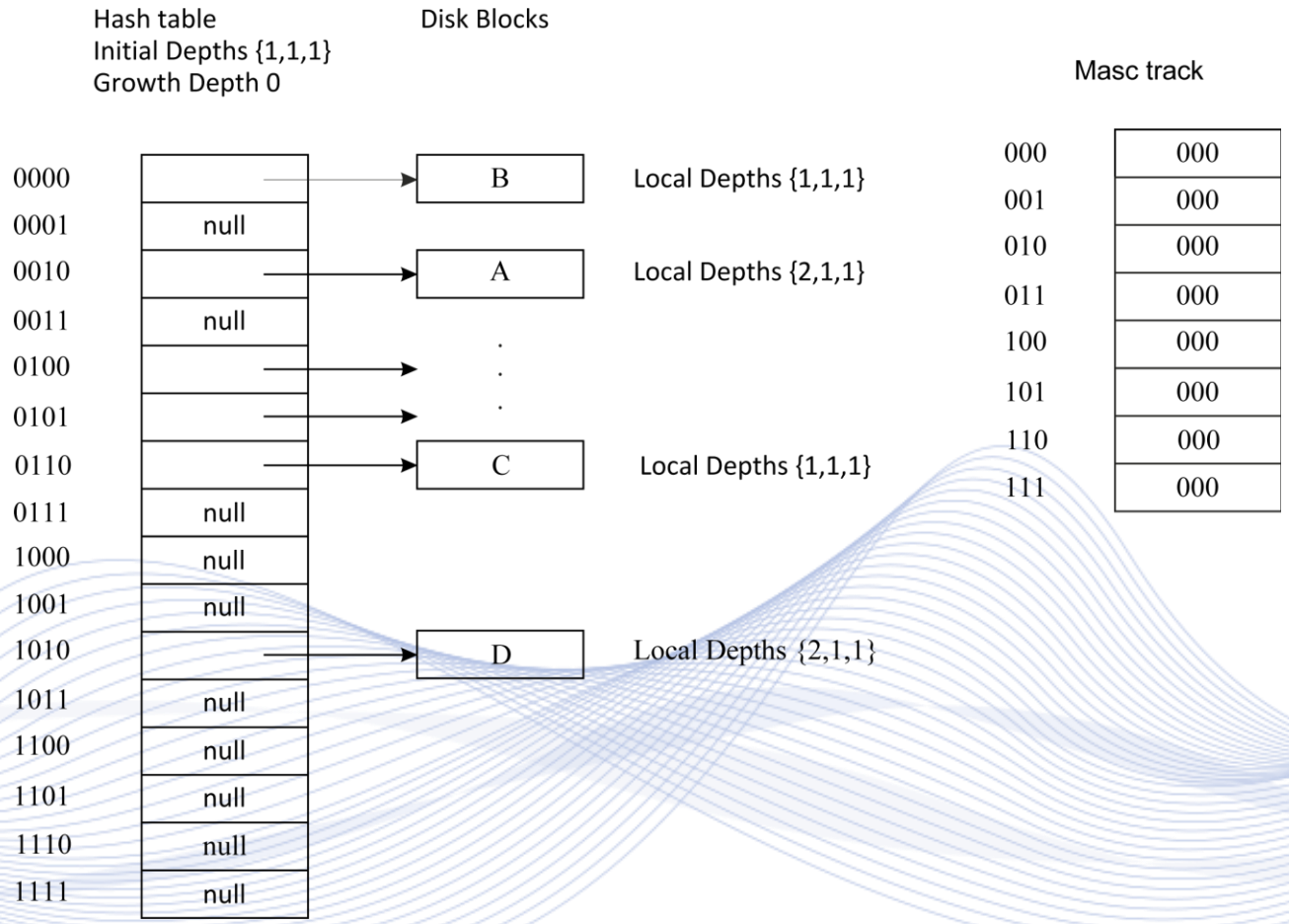
4. Placement of an appropriate number of zeros in the beginning of the resulting address to construct the $(d_1+d_2+d_3+d_g)$-bit hash address. Figures 6-7 show the initial state and the first example of extension respectively [LIN01].

Indexing Techniques



Initial state of the TEMA method hash table.

Indexing Techniques



TEMA method hash table after the first execution.

Indexing Techniques

- A hashing technique based on the dominant image colors has been proposed in [RAV99].

- All colors in an image are mapped to a 25 record color search table. The regions with three dominant colors are indexed using the following equation :

$$Index = C_1 * 25^2 + C_2 * 25 + C_3.$$

Indexing Techniques

- C_1, C_2, C_3 are the values of the most significant colors in the image (in decreasing order of significance).

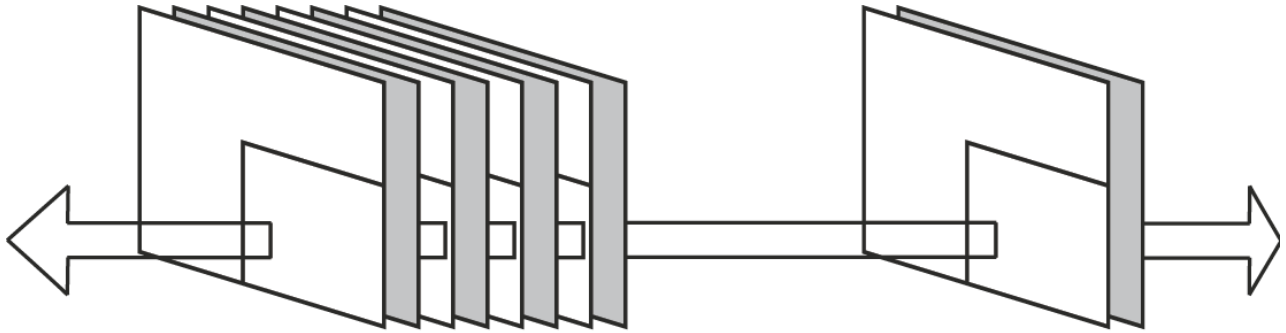
- The relative position of the region and its normalized area is stored with the index (15.7.18). Similar images will be stored next to each other.

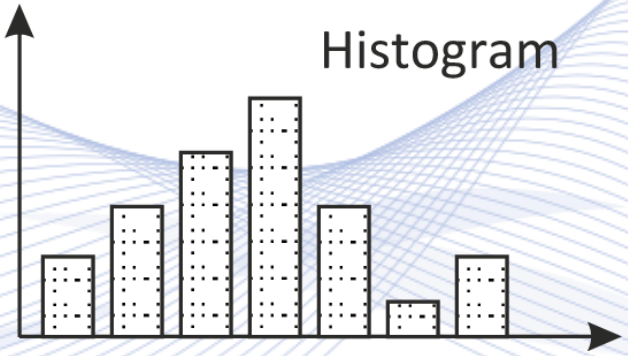
- With the selection of a specific threshold, a number of neighboring images can be retrieved.

Indexing Techniques

- Another indexing technique based on the wavelet transform has been proposed in [Albuz01].
- The vector wavelet transform coefficients of an image are computed.
- For each image, the energy values of the various wavelet bands is stored in the energy vector.
- The next step consists of the calculation of histograms for each element of the energy vector. The problem is shifted to the calculation of the histogram for each wavelet band, based on the total band energy.

Indexing Techniques





Calculation of the energy histogram in a wavelet band.

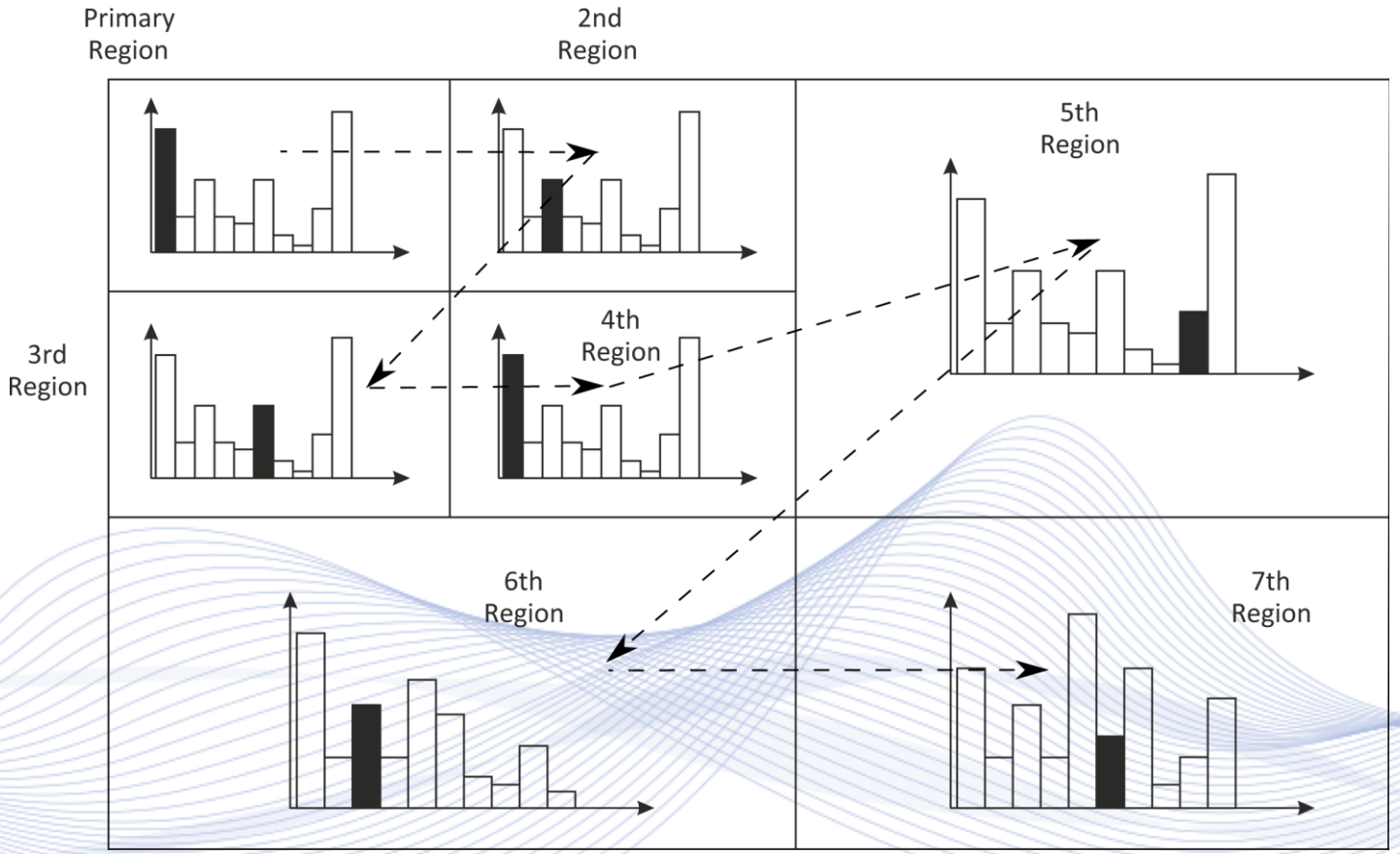
Indexing Techniques

- Each histogram shows the energy distribution of a specific band of all images in the database.

- In the third step, for each image, a number is assigned to each band based on the occupation of positions in the corresponding histogram.

- For each band, these numbers constitute the new feature vector. The band feature vectors are sorted using the Morton scan, starting from the lowest frequency band in the upper left corner to calculate the image key.

Indexing Techniques



Use of the band energy histogram.

Bibliography

[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr