UNIVERSITE DE GENEVE

CENTRE UNIVERSITAIRE
D'INFORMATIQUE
GROUPE VISION

Date: October 1, 2000
N° 00.06

TECHNICAL REPORT

VISION

Content-based *Video* Retrieval: An overview

Stéphane Marchand-Maillet
Viper team
CUI - Université de Genève

Computer Vision Group
Computing Science Center, University of Geneva
24 rue du Général Dufour, CH - 1211 Geneva 4, Switzerland

e-mail: marchand@cui.unige.ch

Contents

Abstract

Content-based Image Retrieval systems (CBIRS) start flourishing on the Web. Their performances are continuously improving and their base principles span a wide range of diversity. Content-based Video Retrieval systems (CBVRS) are less common and seem at a first glance to be a natural extension of CBIRS. In this document, we summarise advances made in the development of CBVRS and analyse their relationship to CBIRS. While doing so, we show that CBVRS are actually not so obvious extensions of CBIRS.

1 Introduction

Content-based Video Retrieval (CBVR) systems appear like a natural extension (or merge) of Content-based Image Retrieval (CBIR) and Content-Based Audio Retrieval systems. However, there are a number of factors that are ignored when dealing with images which should be dealt with when using videos. These factors are primarily related to the temporal information available from a video document. While these factors may complicate the querying system, they also may help in characterising useful information for the querying.

The temporal information firstly induces the concept of motion for the objects present in the document. When in CBIRS, it is the list or organisation of such object which is search for, video retrieval may imply the retrieval of a behaviour of an object throughout the document. Two video documents may therefore contain the same objects but little relevance may be found between the two in this search context. It is therefore essential to encode within the indexing of a video document the behaviour of all objects throughout the document. Another aspect that does not exist in CBIRS and that should be taken into account in CBVRS is the structural organisation of the document . A video document can typically be split into a hierarchical structure (see section 4.3). Another issue in video retrieval is the complexity of the querying systems. A very elaborated retrieval systems would allow flexibility for the user to specify its query parameters. Query-by-example systems typically used in CBIRS require the user to show

the system one or more document similar to what he is looking for. Although this seems natural for images (the whole content of an image may be scanned in one glance), this querying process is far more complicated to adapt into the context of video documents.

2 The need for CBVRS

Before going into details, it is worth reviewing the potential applications of a Content-Based Video Retrieval System. This technology is fairly recent and it is currently necessary to examine where it would just replace existing systems, where it can really bring some improvement and where it will open new possibilities. While doing this, we should review all functionalities a *useful* CBVRS should include.

By definition, a CBVRS aims at assisting a human operator (user) to retrieve a video sequence (target) within a potentially large database. Three major cases may be distinguished.

- The user has a specific sequence in mind and knows it is included within the database. In this case, the target is unique and corresponds to specific criteria. The user will be able to describe fairly precisely the targeted sequence and will be able to see at the first glance whether a suggested sequence corresponds to his wish. Since he knows that the document is in the database, the user will keep querying until he finds the document. In this case, an indexing using text such as keywords or title should suffice.
- The user has a specific video in mind and does not know if the document he is looking for exists in the database. The problem here is to provide a precise search tool so that the user can quickly make a decision as to whether the target is in the database or otherwise.
- The user simply searches for a document by referring to its topic or some event occurring within it. In this case, the search should be hierarchical, in order to guide the user through the search. Relevance feedback is important in this case since it would be the means for user interaction by allowing the user to filter the response of the system. An important issue in this case is the representation under which the response is given. Since the user is not supposed to know the documents, he should be able to scroll all responses quickly for selecting the relevant (or irrelevant ones). It is therefore essential the each document is represented by a short but complete and comprehensive description.

Potential users may arise from different horizons. In the case of CBVRS, main users can be listed as follows.

- News Broadcasting. The need for keeping complete archives induces a large volume of typically short documents. The specificity of news report video documents makes an automated CBVR very attractive to retrieve documents with respect to a specific topic, place or appearance of a given character. One should also note that the audio stream attached to all such documents as well as captions generally present in news reports form important cues for an automated retrieval process.
- Advertising. Here again, documents are typically short. An example of retrieval may concern the fact of retrieving all document w.r.t a style of shooting. Such a characteristic is generally difficult to express using text so that a visual-based approach is highly desirable.
- Music video clips. Here again, the characteristics of such documents may be difficult to express based on given textual annotations stored on a database. For this type of specific applications, one may think of retrieving documents on the basis of some dance step described by *e.g.*, a sketch.
- Distant learning. Two types of educational documents can be distinguished. Firstly, lectures where the main content is given by the audio stream and practical courses when both the visual contents and the audio stream are of importance.
- Video archiving
- Medical applications.

From this analysis it seems clear that there is a need for CBVRS in different domains. The major issues for ensuring the usability of such systems emerge as follows.

- Obtaining a compact and complete video sequence representation
- Providing the user different search strategies adapted to the type of search he is doing.

The next section reviews the internal representation for automatically deriving a representation of a multimedia document.

3 Spatial scene analysis

This section first reviews some visual document processing operations that will be essential for automatically extracting an extensive description of a document. Feature extraction aims at characterising a list of properties (called *feature vector* or *document signature*) for each component (pixel, frame region, frame, sequence) of a video document.

Feature extraction operations rely on the analysis of the human visual system (HVS) and range from simple statistics to elaborated model-based filtering techniques. Another distinction separate unsupervised (*i.e.*, generic) feature extraction operations from supervised (*i.e.*, based on heuristics or training) recognition tasks.

The analysis of elements such as colour and texture aim at characterising features in the spatial space (as opposed to the temporal domain). Experience acquired from CBIR studies may be fully transferred to video in this case. More specific to the video domain is the use of temporal and possibly audio information to characterise further a document. Some key-advances made in these directions are reviewed below.

3.1 Colour feature space

Colour is an important cue for measuring the similarity between visual documents. Historically, colour features have been the first features used in the context of CBIR.

Colour statistics are used for measuring global or local dissimilarities. Global colour features are analysed through histograms. These histograms offer the advantage of being invariant under rotation, translation and many other geometric operations. However, such a global analysis does not allow for characterising the spatial organisation of the colour within the 2D spatial domain. There is therefore a need for refining this technique with using colour layout features.

Features encoding colour organisation within the document are often based on blocks and are either blind (*i.e.*, rigid partitions) or adaptive (*i.e.*, segmentations). A feature vector is attached to each unit part of the spatial domain, and it is the relationship between neighbouring image parts which is encoded as feature.

Whatever their exact definition is, the comparison of all these features calls for the definition of distance measures between colours and colour histograms. Colour distances are defined implicitly through the use of a given colour space. Colour spaces such as YIQ and La*b* were designed based on empirical measurements of the HVS so that the Euclidean distance defined in the respective colour space fairly represents a perceptual distance (see for example [16] for a thorough study).

Colour statistics based on training or more generally supervised learning may allow to distinguish between pixel colours having a specific semantics (*e.g.*, skin colour space [18]). Once histograms have been obtained, they may be compared using histogram distances (see [33] for examples).

3.2 Texture feature space

The analysis of textures requires the definition for a local neighbourhood corresponding to the basic texture pattern. It makes no sense to study the texture of an isolated pixel. Typically, the analysis is done via the mapping of the texture onto the response of one or a bank of pre-defined filters against the image (wavelets, Gabor filters). Once again, this response space aims at allowing the use of a similarity measures between feature vectors. The response image may therefore be analysed using a similar set of tools than for the analysis of colour.

Another approach defines textons [12, 23] as the basic builders for any texture. Each texture is decomposed using these building blocks and the parameters of the local texture are obtained. Typical texture features include orientation and coarseness [12]. In [31], texture models are learned so that geodesic active contours are able to segment texture regions (then considered as uniform or consistent patches), thus extending the classic snake-based segmentation algorithms. Supervised learning of textures

is done via texture samples. One classic texture database is the Broadatz texture album. Other texture databases include VisTEX (MIT), CUReT and the MeasTEX framework (see [24] for pointers). An analysis of colour and texture dissimilarity measures for CBIR is proposed in [33].

3.3 Supervised feature spaces

More complex features may be defined for parsing the contents of a video document. One example of such feature is the development of face detection algorithms. Human faces are widely recognised as a useful cue for video indexing. Here again, the same applies to image indexing but more information can be obtained in the context of video indexing.

The presence of one or more human faces in an image allows for its classification of the image under the category "person", "crowd" or equivalent. The recognition of persons present in an image allow for further classification. The same classification readily applies to video. In a video document, a further step can be made by exploiting the frequency of occurrence of one person for the classification of the document with respect to the ID of that person. Furthermore, face localisation may constitute the starting point of the segmentation of the person within the video document. Motion information obtained through that segmentation will allow further classification.

Face localisation algorithms are mostly based on supervised techniques such as Neural Networks [9] or HMM [27, 26]. Face recognition is then applied by PCA techniques such as EigenFaces [44] or supervised learning such as NN or HMM [30, 39].

The retrieval of text in a video document may also improve its understanding. It is often the case that textual annotations are readily available within the document itself. Using text specificity such as geometrical shape and contrast, captions or credits may be extracted and processed through OCR for completing the document indexing (see *e.g.*, [2]). Finally, combining models for understanding may permit high-level interpretation [32].

4 Temporal analysis

The temporal dimension of a video document contains an information that is specific to this type of document. The temporal analysisof that document typically requires its partitioning into basic elements. It is now recognised that this partitioning can operate at four different levels of granularity.

- Frame level: Each frame is treated separately. There is no (or little) temporal analysis at this level.
- Shot-level: A *shot* is a set of contiguous frames all acquired through a continuous camera recording. The partitioning of the video into shots generally does not refer to any semantic analysis. Only the temporal information is used.
- Scene-level: A *scene* is a set of contiguous shots having a common semantic significance.
- Video-level: The complete video object is treated as a whole.

One key level is the shot-level. Three types of shot boundaries are generally recognised.

- *Cut*: A sharp boundary between shots. This generally implies a peak in the difference between colour or motion histograms corresponding to the two frames surrounding the cut. Cut detection may therefore simply consist in detecting such peaks. Adding any form of temporal smoothing will also improve the robustness of the detection process.
- *Dissolve*: The content of last images of the first shots is continuously mixed with that of the first images of the second shot. The major issue here is to distinguish between dissolve effects and changes induced by global motion. Fade-in and fade-out effects are special cases of dissolve transitions where the first or the second scene, respectively is a dark frame [38].
- *Wipe*: The images of the second shot continuously cover or push out of the display (coming from a given direction) that of the first shot.

While cut detection is relatively easy due to the abrupt nature of the transition, dissolve and wipes are more difficult to detect. Some efficient solutions exploiting the compressed structure of MPEG files have been proposed in [10, 29, 48], based on global motion estimation and segmentation. More elaborated

effects such as mosaics and whirls may be considered as dissolve effects at high scale. However, depending on the particular type of frame mixing technique, dissolve detectors may be mislead by the apparent motion induced by such effects (*e.g.*, whirls). We are not aware of any technique specialised in detecting elaborated gradual effects.

The definition of a scene is based on a deep understanding of the contents of the shots. Automated scene annotation rely on a high-level clustering of shots where the indexing data derived from shots composes feature vectors (see below). Depending on the video, the segmentation of shots may lead to a small (*i.e.*, manageable) set of objects (shot representations). In this case, the definition of scenes can realistically (and reliably) be solved by a human operator. It is important to note that a shot segmentation performed by a human operator may not be fully reliable, due to the fact that this task is tedious and calls for a constant concentration. The development of semi-automated segmentation and annotation tools is therefore important in this context [28].

4.1 Motion feature space

While colour and texture and their organisation characterise the content of a still document, when processing video documents, it is essential to also account for the temporal dimension. The temporal features should provide the information regarding the global (temporal) organisation of a video document. Temporal information is generally translated into a motion characteristic. Motion analysis is made on matching consecutive frames one with another. A (possibly directed) search is made between pixel blocks of two consecutive frames. Statistics allow for characterising global motion (dominant or camera motion) and object motion. Using this information, one can compensate for the global motion leaving only object motion so that temporal information can eventually be used for characterising (*e.g.*, isolating) objects within the document.

It is important to note that motion information is often readily available from the document data itself (*e.g.*, in MPEG compressed data). An efficient motion analysis technique should therefore use this compact representation. This is for example the case in [10, 28, 29, 45].

4.2 Audio feature space

When available, the audio stream attached to a video document may be of great help in understanding the document. This is the case for example in news broadcasts where the audio stream generally corresponds to what is displayed in the frames.

Typically, audio processing techniques are based on the analysis of the energy contained in the audio signal [34]. The signal is divided into *audio frames*, corresponding to few milliseconds of the signal. Features such as Mel-frequency cepstral coefficients (MFCC) and associated statistics are then derived for characterising and classifying the audio frames [49].

One task consists in distinguishing between speech and music or background noise in the audio signal. Algorithms exist that can achieve this task with a good accuracy, based on the fact that speech and music have fairly different spectral distributions and temporal pattern. An example of such technique, based on a connectionist-HMM framework can be found in [46].

Speech transcription and segmentation algorithms are well-advanced and allow for segmenting the video document when synchronising audio and video cuts. Speech summarising techniques are also well-elaborated and help a great deal in the definition of a summary of the video document.

Combining visual and audio streams re-enforces the definition of a generic content-based *multimedia* retrieval system (CBMRS). An example of application exploiting fully the multiple nature of information (audio, textual captions and video) is described in [11, 43].

4.3 Indexing

Once the video segmentation is operated at a desired level, the indexing of the document is performed by creating some meta-data which will be attached to this document for quick reference [17]. The content of the meta-data varies, depending on the application towards which the database is oriented. For generic video documents, this data generally includes video object (shot, scene) boundaries along with some characteristic and visual representation. One common representation is the choice of one or more key frames within the shot or the sequence. Depending the assumptions under which the segmentation has been performed, all frames within a basic shot should normally be consistent with each one another.

Heuristics for choosing one or more key-frames can therefore be derived. The simplest relies on the global position of the frames within the shot (first, middle, end). Some other characteristics such as the corresponding audio stream may also be used for efficient key frame detection [8, 14]. The process of re-segmenting the shots with respect to some heuristics reflecting a comprehension of the video content is referred to as video *micro*-segmentation [25].

5 Similarity measurements

When using feature spaces, one essentially operate a reduction in the dimensionality. This makes typical clustering algorithms more efficient for the grouping of object with respect to these features. The last part of this section briefly lists some clustering algorithms that are used for element grouping.

5.1 Feature clustering

Segmentation calls for the grouping of objects into a (possibly pre-defined) number of clusters. Objects are represented as vectors in a N-dimensional space. Typical clustering algorithms may be used in this context. These techniques generally assume that each cluster has been generated from a PDF of a given (often multi-Gaussian) form. k-means, L.B.G or more generally E.M. algorithms (perceptron-based technique are among them) use this assumptions to optimise the parameters of such a representation. They iteratively associate a label to each vector so that the total repartition matches optimally the initial assumption.

While EM techniques are very popular in image segmentation, hypothesis tests are widely used for speech segmentation. Hypothesis tests are used for the analysis of pairs of vectors. If some prior information is available on the form of the PDF, one may therefore test whether two vectors were generated from the same source or otherwise.

5.2 Distance measures

To measure the similarity between video sequences, one needs to derive a distance which either accounts for the temporal deformation of the video (varying frame rate) or use features which are invariant under temporal distortions. In other words, it seems clear that a frame-to-frame difference measure will not be efficient in comparing two generic video sequences.

In [1], a video distance measure is proposed. The video sequence is represented as a string of symbols taken from a vocabulary and its is the similarity between corresponding strings that is taken as a similarity value for the two video sequences.

Alternatively, much work as been devoted in the comparison of still images in the context of content-based image retrieval (CBIR) (see *e.g.*, [3, 33, 40]). Therefore, one may think of comparing video sequences using such techniques in conjunction with representative frames of the sequence. However, in this case, the temporal information (*e.g.*, motion) would be lost and the retrieval would be done only on 2D visual criteria. Moreover, such a technique assumes the existence of representative frames, which may be acquired within the process of video segmentation, defined next.

6 Video browsing

Going quickly through a video calls for defining a compact representation of that document.

6.1 Slide representation

Most of video displaying techniques are based on the creation of a slide show using selected keyframes attached to the corresponding audio stream part or some textual annotation [11, 14, 22, 21]. Such a slide show can be displayed as a still image mosaic or by re-generated a shorter video sequence, based on few frames surrounding each keyframe.

6.2 3D browsing

The temporal information constitute a third dimension which may be treated as such. This is for example the case in [5] where a VRML-based virtual world is created to browse the video. This space is merely a development of the video key-frames in the temporal direction. This viewing relies on indexing and creates an interactive space using which the (experienced) user may quickly parse the video.

Although it is fairly intuitive to add the temporal development of the video as a third dimension, one should keep in mind that this dimension is not consistent with the 2D (spatial) space created by each frame. This dimension owe therefore a different treatment and possibly a different representation. However, 3D virtual spaces really form a strong alternative to classical video browsing.

7 Video querying

The problem of searching a video document calls for that of formulating a meaningful and clear query. For content-based image retrieval, the system of query-by-example is relatively intuitive since it caters for cases which would be difficult to solve using simple text queries. For video documents however, things are no so straightforward since a query-by-example would require the user to have a video at hand already. One of the major problem is the excessive dimensionality of the search space induced by the temporal information. In order to reduce this dimensionality, different approaches are taken which all introduce advantages and shortcomings.

7.1 Visual query

Since content-based retrieval systems address the search for visual objects, it seems natural to proceed this search based on examples of such visual documents. Different CBIR systems based on a query-by-example (QBE) are available on the WWW, either as commercial products or research prototypes. QBE-based CBIRS can be divided into different groups. In the most basic category, the user is asked to choose one image which is supposed to resemble the one he is searching for. As a result, the system returns in decreasing order the images it finds the most similar to the example given. An enhancement of such a system allows for the user to choose more than one example so that the query combines all common features in these images. The generalisation of that system uses feedback so that the search emulates an on-line learning of the user's expectations. Negative and positive feedback rates are used as equivalent AND and NOT operators. Another refinement of this principle consists in using only parts of the images for the query. This requires the definition of a perceptually-meaningful segmentation technique through which images in the database are pre-processed.

To offer a complete flexibility to the (artist) user, some systems allow for sketching the example image. Again, different levels of such systems exist, ranging from the simple sketch tool to more elaborated tools allowing for the dynamical construction of hierarchical structures (based on logical operators) describing the targeted image. The major drawback of such tools is the complexity induced in the query formulation.

Less examples of QBE-based video search tools exist. One major reason may be the difficulty in describing a video document in a simple and easy-to-represent way (see section 6). Following our earlier analysis, a video example may include either visual still information (*e.g.*, key-frames) or motion information.

7.2 Motion query

We see motion as essentially the only way of representing the temporal information contained in a video document. Motion-based query is therefore an attractive feature of a video search engine. The problem is the formulation of such a query. Motion-based queries can be seen as counter-intuitive in the sense that the user is asked to represent a motion in some still fashion. It seems clear that solving the problem of the formulation of a motion-query is to be made in parallel with that of representing the motion information in the indexing task. Such an approach will facilitate the comparison of the user demand with the available information.

JACOB [7, 6] is a colour- and motion-based video search engine. The user is asked to choose global motion orientation and magnitude within each quadrant of the frames. VideoQ [13] also includes a motion-based search engine. The user is asked to represent a question such as "A red blob moving upwards" through a sketch drawing. In this example, this type of queries may be combined with colour and

keywords. These instances of CBVRS highlight the difficulty in constructing a user-friendly environment for elaborated video queries.

7.3 Textual query

Although it seems intuitively clear that using textual keywords is the simplest way of expressing a query, and that it can be efficiently replaced by the above systems in some cases, we assess this type of query formulation in the context of the above analysis.

QBE-based systems have demonstrated their superior descriptive power, when compared to text-only querying systems. However, it turns out that both querying systems are needed. In other words, textual query seems to be a necessary complement to a QBE-CBIRS. The reason for this is that textual query allows the user to insert some "personalised" data within the query. Using QBE, the user is limited to what he is proposed by the system. Since sketching or composing an image is fairly difficult, the user has to cope with what is available, namely often a set of images drawn randomly from the database (typical CBIRS initialisation). Using keywords, the user may express high level concepts which would be difficult to express through QBE. For images, all is working as if the query was formulated like: "Retrieve a document which contains `[keywords]` like these `[example document]`".

Textual annotation therefore is related to high-level semantic concepts. By definition, these are difficult to create automatically. There is therefore the need for users to perform this task. Since it is fairly tedious, it cannot be done reliably by one human operator. One possibility is to use user judgements to incrementally create this textual annotation database on the basis of visual-only querying. Such a "auto-annotation process" can be done in the manner of collaborative filtering to improve confidence in user judgements.

The same clearly applies to image and video querying systems (in the sense that video database may be searched using images). For video documents, textual querying may be of even more comparative importance since it allows for completing (or replacing) the expression of the motion component of the query.

In summary, textual query is very important since it offers to the user the possibility to complete the weaknesses of the querying interface. The more complete and efficient the interface for visual and motion query becomes, the less textual querying is needed.

7.4 Combining query types

The necessity of using a combined query system appears clearly from the above analysis. Querying systems should typically be organised so as to cater at maximum for all possible users' needs. Also, each type of querying should concentrate on representing the part of the information it has been best at. The pseudo-query thus created should incorporate all search characteristics and overcome the problem of user-dependency of querying.

Combining query types is however not so trivial since it calls for mixing parameters which may not fully be coherent with one another. Different strategies may be envisaged.

One may think at using each type of query separately and combining the different results thus obtained with respect to a common relevance measure. This however defeats the advantage of being able to simulate operations (like AND, OR, NOT) on the desired characteristics. Another simple way to combine the querying system is to normalise the influence of each and to ask the user itself to provide weights for each component of the query. This is not an acceptable solution for two major reasons. Firstly, it complicates the formulation of the query and also, these weights may not be trivial to express so that this solution would simply transfer the problem onto the estimation of such weights.

8 Conclusion

In this document, we briefly reviewed the type of processing operations that are available for automatically characterising the contents of a multimedia document. The mapping from the raw data to some feature spaces chosen with respect to a modelling of the HVS allows for segmenting the document into parts that have a perceptual meaning.

An efficient document search will be based on the (internal and mutual) characteristics of these basic perceptual components. While efficient information can be characterised through these techniques, there

is a strong need for studying interfaces which would allow the user to manipulate this information. This in typical when addressing the problem of video-query where the query formulation is one difficult part of the problem.

References

[1] D. A. Adjeroh, M. C. Lee and I. King. A distance measure for video sequences. *Computer Vision and Image Understanding (special issue on content-based access for image and video libraries)*, 75(1/2):25–45, July/August 1999.

[2] L. Agnithotri and N. Dimitrova. Text detection for video analysis. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 109–113, Fort Collins, Colorado, USA, June 22 1999.

[3] Selim Aksoy and Robert M. Haralick. Textural features for image database retrieval. In *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries in conjunction with CVPR'98*, Santa Barbara, CA, 1998.

[4] M. Yeung andB.-L. Yeo andB. Liu. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, 1998.

[5] S. Vogl andK. Manske and M. Mühlhäuser. A VRML approach to web video browsing. In *Proceedings of the Multimedia Computing and Networking 1999 Conference*, pages 276–285, San Jose, CA, 1999.

[6] E. Ardizzone and M. La Cascia. Automatic video database indexing and retrieval. *Multimedia Tools Applications*, 4:29–56, 1997.

[7] E. Ardizzone, M. La Cascia and D. Molinelli. Motion and color based video indexing and retrieval. In *Int. Conf. on Pattern Recognition (ICPR'96)*, Vienna, Austria, 1996.

[8] Y. S. Avrithis, A. D. Doulamis, N. D. Doulamis, and S. D. Kollias. A stochastic framework for optimal key frame extraction from MPEG video databases. *Computer Vision and Image Understanding (special issue on content-based access for image and video libraries)*, 75(1/2):3–24, July/August 1999.

[9] S. Ben-Yacoub, B. Fasel and J. Luettin. Fast face detection using MLP and FFT. In *Proc. Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, pages 31–36, Washington, MD, 1999.

[10] P. Bouthemy, M. Gelgon and F. Ganansia. A unified approach to shot change detection and camera motion characterisation. Technical Report 1148, IRISA, Rennes, France, 1997.

[11] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young. Automatic content-based retrieval of broadcast news. In *ACM Multimedia 95*, San Francisco, CA, 1995.

[12] K. I. Chang, K. Bowyer and M. Sivagurunath. Evaluation of texture segmentation algorithms. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR99)*, volume 1, pages 294–299, Fort Collins, Colorado, 1999.

[13] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong. VideoQ – an automatic content-based video search system using visual cues. In *ACM Multimedia Conference*, Seattle, WA, 1997.

[14] M. G. Christel, M. A. Smith, C. R. Taylor, and D. B. Winkler. Evolving video skims into useful multimedia abstractions. In *ACM CHI'98 Conference on Human Factors in Computing Systems*, Los Angeles, CA, April 1998.

[15] S. Eickeler and S. Müller. Content-based video indexing of TV broadcast news using Hidden Markov Models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2997–3000,, Phoenix, USA, March 1999.

[16] T. Gevers and A. W. M. Smeulders. A comparative study of several color models for color image invariants retrieval. In *Proceedings of the First International Workshop ID-MMS'96*, pages 17–26, Amsterdam, The netherlands, August 1996.

[17] F. Idris and S. Panchanathan. Review of image and video indexing techniques. *Journal of Visual Communication and Image Representation*, 8:146–166, 1997.

[18] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. In *Proceedings of the Conference on Computer vision and Pattern Recognition (CVPR99)*, volume 1, pages 274–280, Fort Collins, CO, 1999.

[19] V. Kobla and D. Doermann. Video trails: Representing and visualizing structure in video sequences. In *Proceedings of The Fifth ACM International Multimedia Conference (MULTIMEDIA '97)*, pages 335–346, New York/Reading, November 1997. ACM Press/Addison-Wesley.

[20] V. Kobla and D. Doermann. Indexing and retrieval of MPEG compressed video. *Journal of Electronic Imaging*, 7(2):294–307, April 1998.

[21] A. Komlodi and G. Marchionini. Key frame preview techniques for video browsing. In *DL'98: Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 118–125, 1998.

[22] A. Komlodi and L. Slaughter. Visual video browsing interfaces using key frames. In *Proceedings of ACM CHI 98 Conference on Human Factors in Computing Systems*, volume 2 of *Student Posters: Cognition and Perception*, pages 337–338, 1998.

[23] Jitendra Malik, Serge Belongie, Jianbo Shi, and Thomas Leung. Textons, contours and regions: Cue combination in image segmentation. In *Int. Conf. Computer Vision*, Corfu, Greece, 1999.

[24] Stéphane Marchand-Maillet. CVBR page, useful links on content-based video retrieval. URL: http://viper.unige.ch/video/index.html, 1999.

[25] Stéphane Marchand-Maillet and Bernard Mérialdo. Outils stochastiques pour l'indexation vidéo. In *CORESA'99*, Sophia-Antipolis, France, June 1999.

[26] Stéphane Marchand-Maillet and Bernard Mérialdo. Pseudo two-dimensional Hidden Markov Models for face detection in colour images. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington DC, USA, 1999.

[27] Stéphane Marchand-Maillet and Bernard Mérialdo. Stochastic models for face image analysis. In *European Workshop on Content-Based Multimedia Indexing, CBMI'99*, Toulouse, France, October 25-27 1999.

[28] J. Meng and S.-F. Chang. CVEPS – a compressed video editing and parsing system. In *Proceedings of ACM Multimedia 96*, Boston, MA, 1996.

[29] R. Milanese, F. Deguillaume and A. Jacot-Descombes. Video segmentation and camera motion characterization using compressed data. In C.-C. J. Kuo, S.-F. Chang and V. N. Gudivada, editors, *Multimedia Storage and Archiving Systems II*, volume 3229 (SPIE Proceedings), Dallas, TX, 1997.

[30] A. V. Nefian and M. H. Hayes III. Face recognition using an embedded HMM. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington DC, USA, 1999.

[31] N. Paragios and R. Deriche. Geodesic active contours for texture segmentation. Technical Report 3340, ROBOVIS, INRIA, Sophia-Antipolis, France, 1998.

[32] R. W. Picard. A society of models for video and image libraries. *IBM Systems Journal (MIT Media Lab Special Issue)*, 35(3/4):292–312, 1996.

[33] J Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'99)*, 1999.

[34] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[35] V. Roth. Content-based retrieval from digital video. *Image and Vision Computing*, 17(7), 1999.

[36] Yong Rui, Thomas S. Huang and Sharad Mehrotra. Browsing and retrieving video content in a unified framework. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pages 237–240, Austin, TX, 1998.

[37] Yong Rui, Thomas S. Huang and Sharad Mehrotra. Constructing table-of-content for videos. *ACM Journal of Multimedia Systems*, 1998.

[38] R. Ruiloba, P. Joly, Stéphane Marchand-Maillet, and G. Quenot. Towards a standard protocol for the evaluation of temporal video segmentation algorithms. In *European Workshop on Content-Based Multimedia Indexing, CBMI'99*, Toulouse, France, October 25-27 1999.

[39] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceeding of the Second IEEE Workshop on Applications of Computer Vision*, Sarasota, Florida, December 1994.

[40] David McG. Squire, Wolfgang Müller, Henning Müller, and Thierry Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13-14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.

[41] H. Sundaram and S.F. Chang. Efficient video sequence retrieval in large repositories. In *SPIE Storage and Retrieval for Image and Video Databases VII*, San Jose, CA, 1999.

[42] C. Trompler, K. Manske and M. Mühlhäuser. Immersive exploration of video content trees. In *CORESA 99*, Sophia-Antipolis, France, 1999.

[43] S. Tsekeridou and I. Pitas. Audio-visual content analysis for content-based video indexing. In *IEEE Int. Conf. on Multimedia Computing and Systems (ICMCS'99)*, volume I, pages 667–672, Florence, Italy, 7-11 June 1999.

[44] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[45] H. Wang, H. S. Stone and S.-F. Chang. FaceTrack: Tracking and summarizing faces from compressed video. In *SPIE Multimedia Storage and Archiving Systems IV*, 1999.

[46] G. Williams and D. Ellis. Speech/music discrimination based on posterior probability features. In *Proceedings of EuroSpeech*, pages 687–690, Budapest, Hungary, September 1999.

[47] W. Xiong and J. C.-M. Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, 1998.

[48] H. H. Yu and W. Wolf. A hierarchical multiresolution video shot transition detection scheme. *Computer Vision and Image Understanding*, 75(1/2):196–213, 1999.

[49] Tong Zhang and C.-C. Jay Kuo. Content-based classification and retrieval of audio. In *SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII*, San Diego, CA, July 1998.

See also http://viper.unige.ch/video/index.html for the URLs related to some of the above publications.