AUTOMATIC VIDEO SEQUENCE ANALYSIS AND INDEXING

Cataldo Guaragnella[1] and Tiziana D'Orazio[2]
[1] DEE – Electrics and Electronics Department, Politecnico di Bari,
Via E. Orabona, 4 – 70125 – Bari, Italy
[2] CNR – italian National Research Council, ISSIA – Institute of Intelligent Systems for Automation
Via Amendola 166/5 – 70126 – Bari, Italy
dorazio@ba.issia.cnr.it

guaragnella@poliba.it; dorazio@ba.issia.cnr.it

ABSTRACT

In this paper video analysis is addressed by means of IDA, an unsupervised clustering algorithm, to automatically select images taken from a video sequence. The IDA selected images set has been described also in the features vector space to extract motion indication parameters and in intra mode to obtain color vector quantization. The extracted features have been used as video content descriptors and used to classify the sequence types. Preliminary results are presented.

1. INTRODUCTION

Video analysis is one of the most challenging subject in multimedia applications. The goal of a video analysis algorithm should be to "understand" or at least "classify" different video sequences on the basis of the semantic content. This is not an easy task: it requires the development of artificial intelligence (AI) procedures using the sequence motion evolution, the number of object evolving in the scene and their shapes. Any AI procedure should be based on few indicators smartly extracted from the video sequence, i.e. a few parameters able to well describe the image sequence content. At this aim several recent video analysis techniques have been developed in many contiguous framework such as video editing and video segmentation into shots, video retrieval, video semantic description, image/video segmentation, text extraction, video skimming and so on. Video skimming, in particular, tries to exploit similarities between pairs of frames of a given video, to allow video reordering, shot detections and storyboarding. Such procedure find applications even in video coding ([7,8]).

In this paper key-frames have been used to describe synthetically the video semantic content in order to create video indexes to be used in video databases. Key-frames, in facts, well candidate to describe the whole video sequence: a given image in a video sequence represents a single vector of an hyperspace, and the whole video sequence can be described by the dotted curve in this space twisting around a given hyperspace point (the whole video sequence centroid).

The video sequence can thus be summarized by means of few signifying samples in the image vector space (frames) being able to describe the whole video sequence.

The crucial problem in summarizing video sequences lies in the selection of the images of the sequence, able to well describe the whole video time evolution; also the number of frames to be used in describing the whole video might be an information bearing parameter: the higher the number of frames required for video representation, the higher the motion inside the video at hand, at least.

Both the number of frames and a proper choice of frames are required to characterize the whole video content. To obtain this goal ISA algorithm was used, an unsupervised algorithm for vector clustering presented in [6]. This algorithm has been developed looking at the application framework: for video analysis, crucial points are repeatability, robustness, automatism and auto-detection of the features number.

Focus is centered on the analysis of video sequences. A reduced set of key-frames is selected to constitute the skeleton of the image sequence.

The Video Key-frames Codebook (VKC - the video key frames selected to represent the video) is obtained clustering vectors in the image vector space; each vector has M×N components, M×N being the rows × columns pixels of each image.

Few frames of the whole video sequence are extracted by IDA. A post-processing procedure, able to select the few coefficients has been applied: the VKC is clusterized again by IDA to obtain few data vector representing the whole video sequence. The final codebook of vectors representing the VKC is the candidate signature to be used in video content comparison.

We think that such features can be representative of the video at hand, and can concur to give a simple and efficient description of the video sequence in a few coefficients (video summarization), create a video codebook for video coding applications, but also could be used in video retrieval systems for the extraction of given video sequences from very large video data bases ([3,4]).

Furthermore, once selected, the VKC frames can be analyzed in intra-mode to obtain the key farmes color vector quantization, in analogy with [6].

The joint use of the color–space features and the video selected features and motion information has been proposed various applications ([1,2,5]); a synthetic description of color, motion and reference frames of a given video sequence can create a robust signature for every comparison or retrieval application.
Experiments on real videos are carried out. Results of the proposed analysis method seem promising.
The paper is so structured: section 2 describes shortly the IDA clustering procedure; section 3 describes the proposed video analysis system, section 4 the simulation framework and preliminary results. Discussion on results closes the paper.

2. IDA CLUSTERING ALGORITHM

The algorithm uses two concentric loops, the outer one introducing new data centroids (feature), while the inner one used to allow the data structure to converge toward the n-th order best partition of the data structure, n being the number of centroids used to describe the whole data structure.
At the generic step, k, of the outer loop, a new trial centroid is introduced. Its placement is chosen displaced of a parametric quantity, D, from the data distribution centroid, in the direction defined by the vector difference between the true data distribution centroid and the centroid of the (k-1) already defined data features: we maintain that, for the most common data structures, where the data distribution is almost continuous on the whole data set, the defined data feature set should be centered close to the true data structure centroid; stated $\{x_i\}_{i=1,\ldots,N}$ the vector input data set and $\{C_i\}_{j=1,\ldots,M}$, where M << N, the defined features at the M-th external iteration loop, the feature are to be considered representative of the whole vector data structure if the relation (1) holds:

$$\sum_{n=1}^{N} \overline{d_n} \approx \sum_{j=1}^{M} \overline{C_j} \qquad (1)$$

Equation (5) states that the centroid of the M data features detected at the generic iteration of the algorithm present a centroid position in the close neighborhood of the vector data set.
The iterative procedure begins placing a feature of the data set in the data set centroid position. The image partition starts when a new trial feature is generated for the data structure. In this case it is not possible to define where the placement of the feature should be more appropriate, so that the new centroid position is placed randomly in the data space at a given user selected distance D from the data set centroid. Iterations, then, take place to allow the data structure to be split into two regions, each one described by the selected centroids. Iterations split the data set on the basis of a distance norm. Here the Euclidean distance has been used (2).

$$D(j)^2 = (\overline{C_o} - \overline{x_j})^T \cdot (\overline{C_o} - \overline{x_j}) \qquad (2)$$

Even if Mahalanobis metrics often works better, in the general case, it reduces to the Euclidean one for very continuous data structure.
Once the classification of the whole data set has been obtained the new centroid of each detected cluster is computed and a new classification step takes place until the convergence to the best classification is reached.
The stop criterion for the inner loop is given in terms of the least mean squared error (MSE) between the true data set and the obtained classification. When the variation of the MSE with the iteration step is below a given threshold, the iterative refinement of the data feature stops.

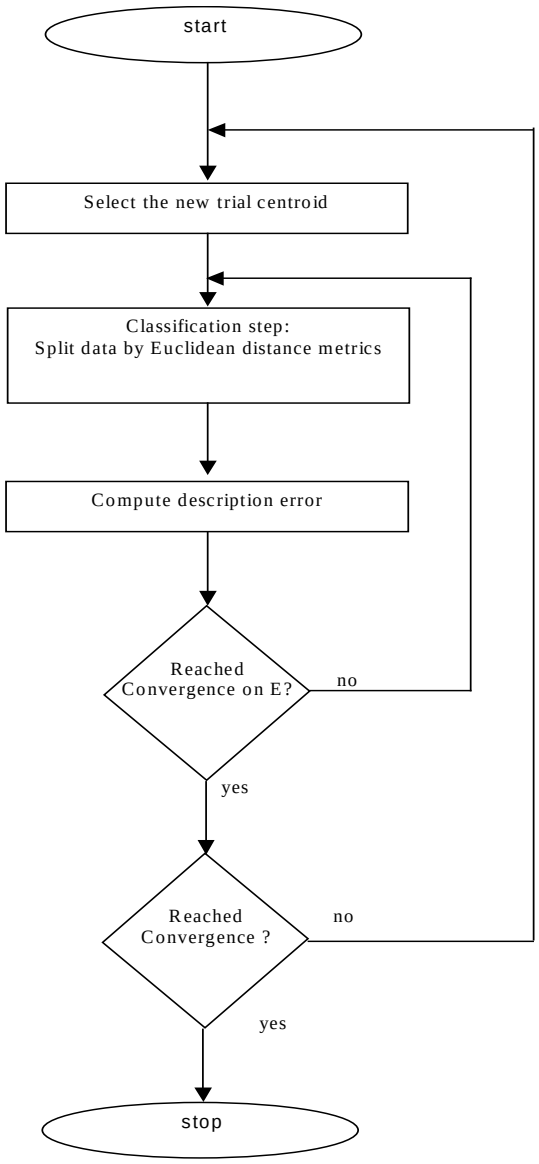

Figure 1: IDA flow chart

At this point a new centroid is placed in the direction of the difference between the data set centroid ($\underline{\mathbf{C}}_o$)and the centroid of the detected features, ($\underline{\mathbf{F}}_o$): a new feature is placed at a distance D from the true centroid in the direction ($\underline{\mathbf{F}}_o$ -$\underline{\mathbf{C}}_o$).
The input distance parameter, D, is passed by the user to the clustering algorithm.
The value to be used for the D parameter depends on very weak knowledge about the data set: when the data set is characterized by several separated features, if the new trial centroid has to be representative of data points in the whole data set, it should fall closer to at least a

few data points to be effective. A too large distance from the given point might cause the event of no data point to be represented by the given trial centroid: in this event, as useless trial centroid are discarded (no data vector represented by the feature – external loop stop criterion), the iterations stop; a too small distance might produce the splitting of an already acquired centroid into two, each representing the same cluster.
For continuous data sets, instead, a too small distance might produce a too populated feature set, so that in a redundant useless data set description.
Figure 1 reports a schematic algorithm block diagram.

3. THE VIDEO ANALYSIS SYSTEM

IDA clustering technique is here adopted to segment the whole video sequence into "clusters of frames": the image space is the hyperspace whose generic vector is an image; each frame in the sequence hence is represented by a vector in this hyperspace.
Figure 2 reports the proposed video sequence analysis system block diagram.
All the images taken from any video sequence, in a given format (say CIF, QCIF, etc.) belong to a limited vector space. If the generic image is considered as a vector of the image space, it presents M×N components (rows × cols) each spanning a limited space, using a color depth of b bits, defined by 2^b (for luminance only images, b=8, and $2^b = 256$ possible colors are commonly used).
The use of IDA has been stressed, in order to reduce the video signature to be used to a very compact set of vectors.
In this work we don't address the well known problem of segmentation of a generic video into shots: literature presents a very wide variety of applications of this kind.

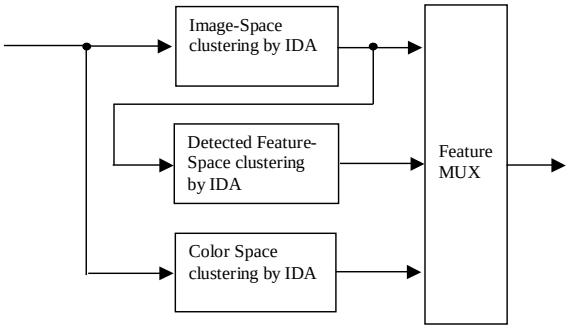

Figure 2: Video sequence analysis system

Here a very challenging goal is pursued: given each shot the whole video has been segmented into (e.g. a video sequence), we try to describe the action present inside each segmented shot by means of a few parameters, in order to give a compact description of "colors, shapes and motion" involved in the video sequence evolution.
A very first attempt toward the definition of a compact set of useful parameters that can be used to describe the video semantic content is presented.
Images of a video sequence span a very narrow subspace of the whole image hyperspace. We'll indicate as a vector of the image space, $\underline{\boldsymbol{x}}$, any given image of the sequence in the stated image space. The video sequence can be considered as a general trajectory in the hyperspace, each point being the generic image at time t. As long as a smooth curve in a given space can be easily described by a set of properly chosen "samples", the whole video sequence at hand can be represented by a (small) subset of well known images.
The problem of describing a video sequence has been addressed also in [9]. When considering the image vector space, the dimensionality of the problem is very large, so that in [9] a preprocessing was introduced to simplify the image description. Preprocessing might be used to simplify the each image content to a few features, thus making easier the clustering procedure in the reduced feature vector space.
Here the clustering procedure is applied to the images of the video sequence without intra-frame preprocessing of the images. The application of the IDA algorithm to the frames of the video sequence allows a better discrimination between frames and reduces the computational load of the algorithm. Furthermore the IDA algorithm reveals more robust, deterministic in obtained results and faster in convergence than the NUSD unsupervised approach proposed in [9].
After the initial video key-frames extraction, the subsequent application of the IDA algorithm to the few determined key-frames is carried out. This time, stated K the number of extracted key-frames of the sequence, for each image pixel a vector is created, its components being the K key-frames corresponding pixels.
The IDA application on this newly defined space allows the determination of the cluster of moving pixels: all the pixels of the key-frames not experiencing motion, assume the same value on all the key frames, while moving pixel present vector components generally different. The application of IDA in this new space allows the highlighting of motion, so that the obtained features, together with the selected key frames, can synthetically describe the semantic image content, including low cost motion indication.
The color information can be useful to complete the whole sequence content, the intra-frame clustering of the key-frame images can be used to give a vector color description, as done in [9].

Figure 3: color regions detected

4. THE EXPERIMENTAL SETUP AND RESULTS

The proposed procedure has been applied on real video sequences. Here results for Akiyo in QCIF-YUV (144×176) color format are presented.
Figure 3 presents color regions detected by the proposed clustering scheme, while fig. 4 presents the automatically selected (grayscale) key-frames of the sequence.

To show the motion indication properties of the clustering obtained in the key-frame vector space, the difference between a couple of classified images is shown in figure 5: the main difference is present on the

Figure 4: Akiyo obtained key-frames

moving border regions of speaker acting in the scene.

5. DISCUSSION

The creation of a Video Key-frame Codebook is introduced to synthetically describe video sequences. IDA unsupervised algorithm has been used at this aim. Each selected group of image of the VKC has been described in intra mode to generate a signifying color signature and a compact description of the video sequence has been constructed, formed by the VKC, motion indicators and image color prototypes.

The assessment of the proposed measure is the goal of the work: here first experimental results are presented.

To compare the effectiveness of the proposed video description technique, 6 different video sequences in QCIF-YUV color format (300 frames each) have been processed and video description indexes have been created (M&D, Carphone Foreman, Akiyo, Salesman, Trevor).

All the video indexes have been stored and used simply as a low populated video retrieving system.

The video sequence Akiyo has been slightly modified in colors with a trivial procedure, and also a different action has been simulated in the video sequence by changing the sequence reproduction frame index with a slowly time changing random function. This modified version of the sequence has been used as a video comparison test.

Several tests have been carried out creating different video sequences; the overall estimation produced on the average fair results, anyway sometimes wrong results occurred, depending on the amount of changes in colors and motion indexes. The wrong behavior of the proposed technique is probably due to the use of the Euclidean norm for the feature comparison: several different indexes have been grouped in a single image signature, so that to obtain better performances, at least a weighted distance metric should be used. The suitable definition of a metric to address video retrieving in this contest is, at the present moment, still an open problem.

Future developments will focus on this subject; more robust indexes will also be added to the whole video content description to enhance selectivity in video comparison, such as texture indexes.

Figure 5: an example of motion indication obtained by clustering key-frames in the key-frame hyperspace

References

[1] P. Bouthemy, E. Francois, Motion segmentation and qualitative dynamic scene analysis from an image sequence, Int. J. Computer Vision, 10, 2, 159-182, 1993

[2] C. Tomasi, , and T. Kanade, Shape and Motion without Depth, ICCV 90, Osaka, Japan.

[3] A. Celentano, E. Di Sciascio, Similarity Evaluation in Image Retrieval Using the Hough Transform, Journal of Computing and Information Technology, Vol.4, No.3, 1996.

[4] K. Hirata and T. Kato, Query by visual example– Content-based image retrieval, in Proc. EDBT'92, Lecture Notes in Computer Science, Springer-Verlag, Berlin/New York, 1992.

[5] P. Salembier, F. Marques, Region based representation of image and video: Segmentation tool for multimedia services, IEEE Trans. Circuits and systems for Video Technology, invited paper, vol. 9 no. 8, dec. 1999

[6] T. D'Orazio, C. Guaragnella, IDA–Iterative Data Analisys applied to Color Vector Quantization, IEEE Proc. ISCCSP 2004, Intl. Sym. On Control, Communications and Signal Processing, Hammamet, Tunisia, March, 21-24, 2004

[7] E. Di Lecce, C. Guaragnella, Personal Mobile Video Communication based on String Image Description, IEEE International Symposium on Signal Processing and Applications, July 1-4, 2003, Paris, France

[8] G.Acciani, D.Girimonte, C.Guaragnella, Extension of the forward-backward motion compensation scheme for MPEG coded sequences: a sub-space approach, IEEE Proc. DSP 2002, 14th Itnl. Conference on Digital Signal Processing, Special Session on Perceptual Image and Video Coding, July, 1-3, 2002, Santorini, GR

[9] G.Acciani, E.Chiarantoni, D.Girimonte, C.Guaragnella, Unsupervised Neural Network approach for Efficient Video Description, Itnl. Conference on Artificial Neural Networks, Madrid, August, 27-30, 2002