Realistic Video Animation for Negligible-Bit-Rate applications

Cataldo Guaragnella
DEE – Electrics and Electronics Department
Politecnico di Bari, Via E. Orabona, 4
70125 – Bari, Italy
guaragnella@poliba.it

Abstract – In personal video communications, the postures assumed by the speaker during the connection are very similar. The speech of the person is strongly dependent on the assumed mouth positions. A convincing synthetic video animation can be obtained using precoded frames taken from the sequence, exploiting the redundancy between the mouth postures and the speech signal to "drive" a synthetic video of the speaker with negligible information added to the speech coding. This application can be considered as a simple realistic video flywheel to be used in very common congestion situation in IP based personal video communication applications. Convincing actions of the speaker to the receiving end user have been produced, basing on a video animation system. Preliminary results are presented.

Keywords – Video animation, wireless, perceptual video coding, audio/video correlation

1. Introduction

In personal IP based video communication applications the problem to face with is the definition of the coding/decoding algorithm able in dealing with the available bit rate on the given communication channel. The high required time variability of the bit rate of MPEG based codecs, even with the use of MPEG-4 FGS encoding, may crash into the rapidly changing available bandwidth on IP based network applications. MPEG-4 codecs exploit the temporal redundancy of several images in a sequence to greatly reduce the required information to be efficiently transmitted to the receiving end. Its fortune depends on the object oriented video coding structure, allowing several bit streams to be multiplexed together to form a single audio/video frame structure to be transmitted. Each of the separate streams contains information about a single Video Object (VO), defined by means of arbitrary shape, motion and residual error coding.

Even for non real-time streaming applications, anyway, the time varying available bandwidth can cause freezes of images and/or degradation of the received image quality.

MPEG-4 can avail on the multi-reference frame predictive scheme: several frames (previous or future, [1,2]) can be considered in order to obtain better predictions of the current frame. This approach requires the complication of the motion estimation hardware at the coding end to obtain several motion fields, and add side information to the coded bit stream to describe coefficients of the polynomial motion fields at the receiving end.

The use of several reference frames in the prediction reconstruction at the receiving end is a powerful method, and can be addressed relaxing the complication of the coder structure for the image time evolution description, when the bandwidth shrinks for long periods, allowing a flywheel for the communication system, to avoid freezes.

Video coding over wireless networks asks for computationally light procedures, able to achieve high compactness of the bandwidth and high image quality. In this case, a good perceived image quality is much more important than the proper coding of the true video evolution: if the perceived video is considered likely by the end user, the semantic content of the video can be assumed received even if the acting speaker at the receiving end of the video communication link is the result of a locally generated realistic video animation of the speaker.

Several authors ([3, 4]) have addressed the video animation problem, anyway the proposed techniques still reveal too computationally expensive to be used in wireless coding applications. Instead, a simpler approach is proposed. In this work, a low cost video animation system is described. The video to be coded is assumed as a video sequence; an appropriate, reduced in size, set of key frames is properly selected to constitute the skeleton of the image sequence.

A video Key-frame Codebook (VKC) is created; the key frame selection is performed on a training video sequence (TVS) by means of an unsupervised efficient approach (IDA, [5]); audio-video correlation is exploited to be used when no video bandwidth is available. Once the VKC has been created, video reproduction can take place: here pre-coded trajectories in the hyperspace, connecting in pairs the selected key-frames, have been used.

The reference frames are connected one to the others and correlated to the speech characterizing the frames in the neighborhood of each selected key-frame. A audio-driven video animation could at least be granted

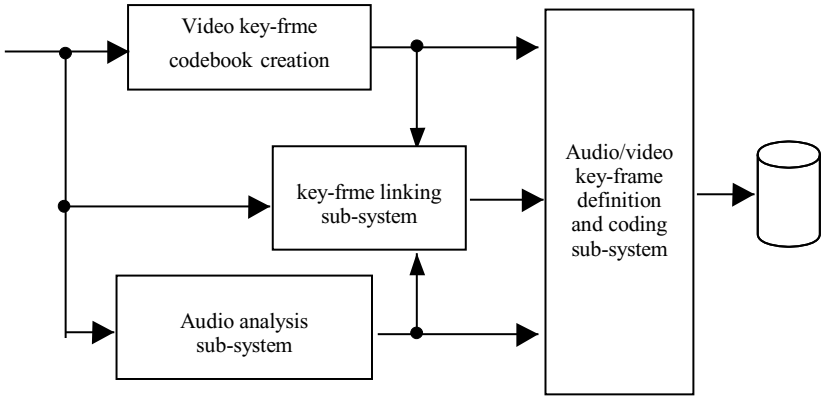


Figure 1: Audio/Video analysis system or the Audio video key frame definition

at the receiving end of the communication link.
The video animation can be obtained as in a finite states system: links between couples of key-frames can be extracted from the TVS, and pre-coded.
To reproduce the synthetic video, only the index to the next key-frame to be reached in the sequence and the number of frames to be synthesized to produce a "natural" evolution of the synthetic video animation are required as information coding.
The paper is so structured: section 2 describes the video sequence space and its use for video analysis/synthesis application; Section 3 briefly describes the unsupervised approach for key-frames extraction and presents the video animation system architecture is presented; section 4 addresses the proposed video animation system to standard sequences and preliminary results. Conclusions and future work close the paper.

2. The video sequence sub-space

The VKC creation consists in the storyboarding of the recorded sequence: as long as the video sequence contains very similar images, only a few of the images of the whole video are required to describe it, at least semantically.
Once the images of the sequence have been chosen as key-frames to represent whole video, any other frame can be someway related to the selected key-frames, to reproduce the true images along the video sequence.
A clustering technique is here adopted to segment the whole video sequence into "clusters of frames": the image space approach is introduced to describe video sequences.
All the images taken from any video sequence, in a given format (say CIF, QCIF or other), 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 (for luminance only images, b=8, and 2b = 256 possible colors are commonly used). Images of a sequence span a very narrow subspace of the whole image hyperspace. Let $\underline{\mathbf{x}}$ be a vector in the image space representing any given image of the sequence. The video sequence can be considered as a general trajectory in the space, 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 subset of images. Any other image along the given sequence can be "interpolated" from the samples in the given space.
A given vector $\underline{\mathbf{x}}$, representing the image of the sequence at time t in the sequence vector subspace, can be related to the "samples" of the space and could be interpolated by means of a proper algorithm, requiring only the knowledge of the subspace position of the image $\underline{\mathbf{x}}$.
The knowledge of the estimated image of the sequence (prediction), obtained by the use of the defined VKC, requires no additional information to be transmitted over the channel (that means lower bandwidth) and no motion estimation phase at the coder end.
The heart of the proposed video coding system is then the proper creation of a VKC for each segmented sequence of the video, here addressed by an unsupervised NN approach, hereafter described.

3. Key-frame selection: the video animation system

For the generic video sequence, we can assume that a starting video sequence can be considered acquired at the coder end, and the key-frames can be extracted from that video sequence. This procedure can be done by means of an off-line initialization procedure, to be performed once at the beginning of the connection, as described in [6].
A training video sequence can be assumed acquired and pre-processed, with the speaker giving a predefined talk to extract audio/video redundancies, in order to select video key-frames and their associated sound information. We refer to this phase as the initialization phase, and the recorded sequence is referred to as the training video sequence. The sequence can be analyzed either on place, for laptop video communications

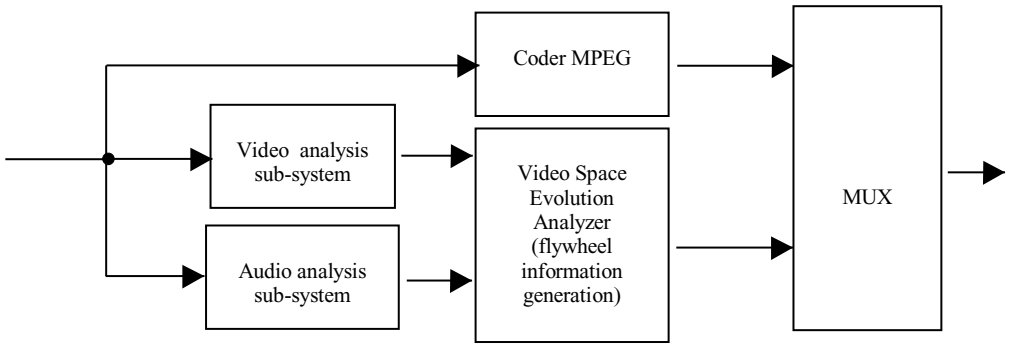


Figure 2: The MPEG video coding generalization scheme using the proposed video animation system

applications, or transmitted to the core network interface, where processing resources are made available for the user, in cases of reduced processing capabilities of the personal communication systems.

The goal of the training video sequence (TVS) preprocessing is to analyze the video sequence in order to:

- select the mostly addressed positions assumed by the speaker along the video sequence, and their associated audio information;
- analyze the whole TVS to find paths linking the selected key-frames in a "natural" video evolution, eventually selecting several paths linking two different key-frames of the VKC defined at the first step.

Once the video animation environment has been initialized, the images and their associated audio information can be transferred to the receiving link-end and used as a tool to synthesize video in case of need when, due to band shrinking, the empty receiving buffer situation is encountered, as described later.

In this way, the coding information required for the video animation system can be reduced to a few bits: to create a likely animation of the speaker, assuming that no freeze of the audio information happens, only the "next key-frame" information of the video sequence should be transmitted, and the "speed" at which it should be reached (number of frames that should be reproduced), together with the selected path to be chosen, where path redundancy is present.

3.1 Key frame selection

The data vector to be used in this phase are the frames of the TVS: frames are arranged in a data vector by row scanning the frame and used in the clustering procedure without any pre-processing.

To select the video key-frames from the TVS, the IDA algorithm is used. This unsupervised clustering algorithm (described in [5]) uses two concentric loops.

It considers a frame as a vector in the sequence hyperspace. The clustering procedure classifies all the frames in the training sequence into few classes, each described by a centroid vector. The iterative procedure starts placing a feature of the data set in the data set centroid position. At the generic step, k, of the outer loop, a new trial centroid is introduced. It 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 ones. The image partition starts when a new trial feature is generated for the data structure. Some iterations take place to allow the data structure to be split into separated regions, each one described by the selected centroids. The data set is split on the basis of a distance norm.

Here the Euclidean distance has been used. 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. When changes in the classifications become negligible, the inner loop iterations stop. The outer loop (introduction of a new trial centroid) is stopped when the algorithm verifies that a higher complication of the data set description is unnecessary: if the variation of the error is below a given threshold, iterations stop and the selected features (data distribution centroids computed in the classification phase) are considered able to describe the whole data set.

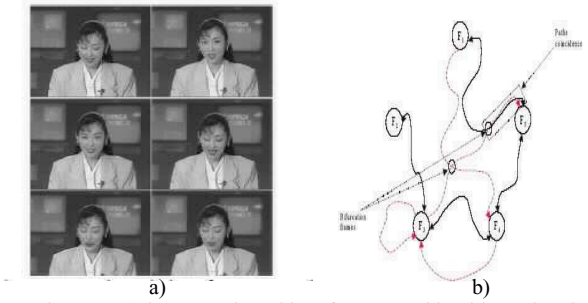
The obtained centroids in the clustering procedure result as a weighted average of the frames in the sequence, corresponding not necessarily with a frame. Key frames of the video sequence are chosen as the frames of the sequence whose Euclidean distance from each selected centroid is minimized.

3.2 The video animation system

The video animation system is here briefly described; both the VKC creation and its use have been pictorially described in figures 1 and 2.

The core function in the video animation tool is the audio feature extraction system whose goal is to extract the information to select the video key-frame number representing the video sequence "state vector" at the given time instant and its subsequent key-frame to be reached in a given number of steps. The Audio/Video correlation exploitation is performed by the audio/video evolution subsystem. After audio and video features

have been analyzed, their features are sent to the video evolution analyzer that maps the audio video evolution parameters in input into the additional fly-wheeling video coding information.
To generate a video animation in likely accordance



a) b)

Figure 3: a) the IDA selected key-frames used in video animation; b)The key-frames link construction: black lines refer to TVS, red lines the generic sequence time evolution

with the input video sequence, only the evolution of the video sequence in the hyperspace has to be known and the speed with which the key frames in the sequence should be reached along the video time evolution.
To create the video sequence, the key frames should be linked by known paths in the hyperspace. Paths have been initially extracted from the video sequence by selecting those portion of frames connecting key-frames in the TVS; by observing the video evolution, several new paths have been created by extracting new sequence pieces connecting different frames.
In this application we have considered paths as bidirectional, so that if there is a path between the i-th keyframe and the j-th one, it is assumed that a likely connection in the opposite direction can be the same path reversed.
Figure 3-b) pictorially represents the obtained paths from the TVS (solid paths) and in dotted the generic video sequence time evolution in the hyperspace. The highlighting of bifurcation frames: points in the hyperspace where the video sequence pass while connecting different couples of frames. The highlighting of bifurcation frames has been obtained, besides the IDA frame classification step, when change in the frame index is required in the video sequence.
The bifurcation frames have been used to connect every couple of key-frames, so that a preliminary video animation system has been finally defined.

4. Experimental results

Preliminary results of video animation have been addressed: no audio/video redundancy exploitation has been yet implemented. Here results refer to the video animation of a sequence on the basis of the selection of the video coding fly-wheeling parameters: the video-only codebook has been created for the standard QCIF-YUV Akiyo video sequence. Video animation preliminary results here addressed can be found at http://www-dee.poliba.it/dee-web/Personale/guaragnella/web-page-video%20animation.htm.
The obtained sequence synthesis has been made selecting randomly the key-frame index and the speed from couples of frames. The comparison between two sequences is shown: the true sequence Vs the synthesized one (same length). Figure 3-a) reports the IDA automatically selected key frames of the Akiyo video sequence. It should be noted that, once the video animation system has been transferred at the receiving end of the communication link, three numbers, at most, are provided to for the animation system the reproduction of a given piece of video: next key-frame number, speed to be used in reproduction of the link path and (eventually) the selected path to be reproduced, in case of redundant paths.

5. Conclusions and future work

A simple video animation system to be used in IP based video coding systems to reduce the freezes problems in video reproduction, due to sudden band shrinking, has been presented.
The good obtained synthetic video appearance seems to promise nice results, even if a fair evaluation of the proposed method should be assessed in the talk driven application. Preliminary results to video-only animation are presented.
Future work will address three main items: the definition of a run time environment able to extract, at the transmitting link end, new paths suitable to be inserted in the video animation environment; a highlighting system to detect bifurcation frames in different links, to ease the linking procedure between several key-frames; audio coding feature extraction and to their connection to the video synthesis system.

References

[1] B. Girod, Efficiency Analysis of Multihypothesis Motion-Compensated Prediction for Video Coding, IEEE Trans. on Image Processing, VOL. 9, NO. 2, Feb. 2000
[2] M. Flierl, B. Girod, Generalized B Pictures and the Draft H.264/AVC Video-Compression Standard, IEEE Trans. on Cyrcuits and Systems for Video Technology, vol 13, NO 7, July 2003
[3] T. Ezzat, T. Poggio, Videorealistic Talking Faces: a morphing approach, Proc. of the AVSP'97Workshop, Rhodes, Greece, Sept.26-27, 1997 – ESCA ISSN # 1018 4554
[4] E. S. Chuang, H. Deshpande, C. Bregler, Facial expression space learning, Proc. of the 10 th Pacific Conference on Computer Graphics and Applications (PG'02), 0-7695-1784-6/02 $17.00 © 2002 IEEE
[5] T. D'Orazio, C. Guaragnella, IDA-Iterative Data Analisys applied to Color Vector Quantization, Submitted to ISCCSP 2004, Intl. Sym. On Control, Communications and Signal Processing, Hammamet, Tunisia, March, 21-24, 2004
[6] 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
[7] 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