New low complexity DCT based video compression method

Tarek Ouni
National Engineering School of Sfax
Road Sokkra, Km 3 Sfax, Tunisia
Email: tarek.ouni@gmail.com

Walid Ayedi
National Engineering School of Sfax
Road Sokkra, Km 3 Sfax, Tunisia
Email: ayedi.walid@gmail.com

Mohamed Abid
National Engineering School of Sfax
Road Sokkra, Km 3 Sfax, Tunisia
Email: mohamed.abid@enis.rnu.tn

***Abstract*—Generally, video signal has high temporal redundancies due to the high correlation between successive frames. Actually, this redundancy has not been exploited enough by current video compression technics. In this paper, we present a new video compression approach which tends to hard exploit the pertinent temporal redundancy in the video frames to improve compression efficiency with minimum processing complexity. It consists on a 3D to 2D transformation of the video frames that allows exploring the temporal redundancy of the video using 2D transforms and avoiding the computationally demanding motion compensation step. This transformation turns the spatial-temporal correlation of the video into high spatial correlation. Indeed, this technique transforms each group of pictures to one picture eventually with high spatial correlation. Thus, the decorrelation of the resulting pictures by the DCT makes efficient energy compaction, and therefore produces a high video compression ratio. Many experimental tests had been conducted to prove the method efficiency especially in high bit rate and with slow motion video. The proposed method seems to be well suitable for video surveillance applications and for embedded video compression systems.**

I. Introduction

The objective of video coding in most video applications is to reduce the amount of video data for storing or transmission purposes without affecting the visual quality. The desired video performances depend on applications requirements, in terms of quality, disks capacity and bandwidth. For portable digital video applications, highly-integrated real-time video compression and decompression solutions are more and more required. Actually, motion estimation based encoders are the most widely used in video compression. Such encoders exploits inter frame correlation to provide more efficient compression.

However, Motion estimation process is computationally intensive; its real time implementation is difficult and costly [1][2]. This is why motion-based video coding standard MPEG was primarily developed for stored video applications, where the encoding process is typically carried out off-line on powerful computers. So it is less appropriate to be implemented as a real-time compression process for a portable recording or communication device (video surveillance camera and fully digital video cameras). In these applications, efficient low cost/complexity implementation is the most critical issue. Thus, researches turned towards the design of new coders more adapted to new video applications requirements. This led some researchers to look for the exploitation of 3D transforms in order to exploit temporal redundancy. Coder based on 3D transform produces video compression ratio which is close to the motion estimation based coding one with less complex processing [3][4][5][6]. The 3d transform based video compression methods treat the redundancies in the 3D video signal in the same way, which can reduce the efficiency of these methods as pixel's values variation in spatial or temporal dimensions is not uniform and so, redundancy has not the same pertinence. Often the temporal redundancies are more relevant than spatial one [3]. It is possible to achieve more efficient compression by exploiting more and more the redundancies in the temporal domain; this is the basic purpose of the proposed method. The proposed method consists on projecting temporal redundancy of each group of pictures into spatial domain to be combined with spatial redundancy in one representation with high spatial correlation. The obtained representation will be compressed as still image with JPEG coder. The rest of paper is organized as follows:
Section 2 gives an overview of basics of DCT based video compression methods. In section 3 we review the basics of the proposed method and the extensions chosen to improve the compression ratio. This is the main contribution of this work. Section 4 presents the results of a comparative study between compression standards with the proposed method. Section 5 gives some analysis and comments about the method. The last section concludes this paper with a short summary.

II. DCT based coding methods

The transform coding developed more than two decades ago, has proven to be a very effective video coding method, especially in spatial domain. Today, it forms the basis of almost all video coding standards. The most common transform based intraframe video coders use the DCT which is very close to JPEG. The video version is called M-JPEG, wherein the "M" can be thought of as standing for "motion". The input frame is first segmented into N N blocks. A unitary space-frequency transform is applied to each block to produce an N X N block of transform (spectral) coefficients that are then suitably quantized and coded. The main goal of the transform is to decorrelate the pixels of the input block. This is achieved by redistributing the energy of the pixels and concentrating most of it in a small set of transform coefficients. This is known as

Energy compaction. Compression comes about from two main mechanisms. First, low-energy coefficients can be discarded with minimum impact on the reconstruction quality. Second, the HVS [1] has differing sensitivity to different frequencies. Thus, the retained coefficients can be quantized according to their visual importance. Actually, the DCT, witch will be used in our video compression approach, is widely used in most modern image/video compression algorithms in the spatial domain (MJPEG, MPEG). Although its efficiency, it produces some undesirable effects; in fact, when compression factors are pushed to the limit, three types of artifacts start to occur: "graininess" due to coarse quantization of some coefficients, "blurring" due to the truncation of high-frequency coefficients, and "blocking artifacts," which refer to artificial discontinuities appearing at the borders of neighboring blocks due to independent processing of each block.[7]

Moreover, the DCT can be also used in the temporal domain: In fact, the simplest way to extend intraframe image coding methods to interframe video coding is to consider 3-D waveform coding. The 2D-DCT has the potential of easy extension into the third dimension, i.e. 3D-DCT. It includes the time as third dimension into the transformation and energy compaction process [3][4][5][6]. In 3-D transform coding based on the DCT, the video is first divided into blocks of M N K pixels (M; N; K denote the horizontal, vertical, and temporal dimensions, respectively).
A 3-D DCT is then applied to each block, followed by quantization and symbol encoding, as illustrated in Figure 1.

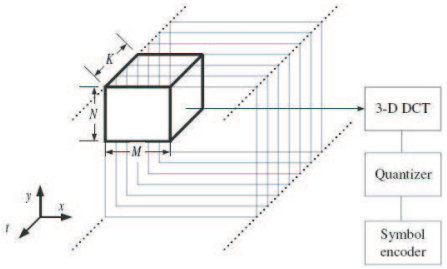

Fig. 1. 3D DCT video compression

A 3-D coding method has the advantage that it does not require the computationally intensive process of motion estimation. However, it presents some disadvantages; it requires K frame memories both at the encoder and decoder to buffer the frames. In addition to this storage requirement, the buffering process limits the use of this method in real-time applications because encoding=decoding cannot begin until all of the next K frames are available. Moreover, the 3D DCT based video compression method produce some side effects in low bit rates, for example the effect of transparency produced by the DCT 3D [8]. This artifact is illustrated by figure 2.

The technics of transformed 3Ds was revealed since the 90s, but the research in video compression was oriented towards the coding based on motion estimation. The design tendency

[1]Human Vision System

Fig. 2. Transparency effect in 3D DCT

of new coding diagrams led some researchers restarting the exploitation of transformed 3Ds in video compression. The coders based on this type of transformation produce high compression ratio with lower complexity compared to motion compensated coding.
3D DCT based video compression methods treat video as a succession of 3D blocks or video cubes, in order to exploit the DCT properties in both spatial and temporal dimensions. The proposed coding method will be based on the same vision. The main difference is how to exploit temporal and spatial redundancies. Indeed, the proposed method puts in priority the exploitation of temporal redundancy, which is more important than the spatial one. The latter assumption will be exploited to make a new representation of original video samples with very high correlation. The new representation should be more appropriate for compression.
Detailed approach description will be presented in the next section.

III. Proposed approach

The basic idea is to represent video data with high correlated form . Thus, we have to exploit both temporal and spatial redundancies en video signal. The input of our encoder is so-called video cube, which is made up of a number of frames. This cube will be decomposed into temporal frames which will be gathered into one frame (2 dimensions). The final step consists of coding the obtained frame. In the following, we detail the method design steps.

A. Hypothesis

Many experiences had proved that the variation of the 3D video signal is much less in the temporal dimension than the spatial one. Thus, pixels, in 3D video signal, are more correlated in temporal domain than in spatial one [3]; this could be traduced by the following expression: for one reference pixel I(x,y,t) where:

- I: pixel intensity value
- x, y: space coordinate of the pixel
- t: time (video instance)

we could have generally:

$$I(x,y,t) - I(x,y,t+1) < I(x,y,t) - I(x+1,y,t) \quad (1)$$

This assumption will be the basis of the proposed method where we will try to put pixels - which have a very high temporal correlation - in spatial adjacency .

B. "Accordion" based representation

To exploit this succeeding assumption, we start by carrying out a temporal decomposition of the 3D video signal, the figure 3 shows temporal and spatial decomposition of one 8X8X8 video cube:

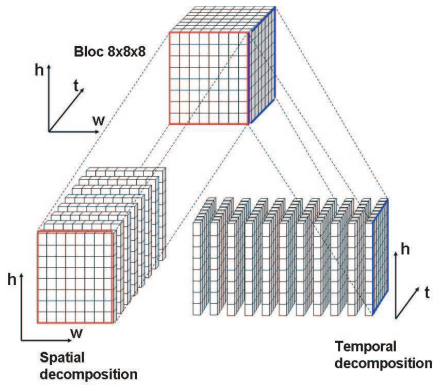

Fig. 3. Spatial and temporal decomposition principle

"Frames" obtained following the temporal decomposition will be called "temporal frames". These latter are formed by gathering the video cube pixels which have the same column rank. According to the mentioned assumption, these frames have a stronger correlation compared to spatial frames. to increase correlation in Accordion Representation we reverse the direction of event frames. Figure 4 illustrates the principle of this representation.

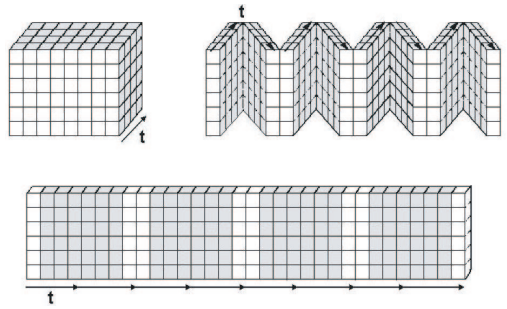

Fig. 4. ACCORDION Representation

Thus, the "Accordion representation" is obtained as following: first, we start by carrying out a temporal decomposition of the video 3D. Then, the event temporal frames will be turned over horizontally (Mirror effect). The last step consists of frames successive projecting on a 2D plan further called "$IACC$" frame.
The "Accordion representation" tends to put in spatial adjacency the pixels having the same coordinates in the different frames of the video cube. This representation transform temporal correlation in the 3D original video source into a high spatial correlation in the 2D representation ("$IACC$"). The goal of turning over horizontally the event temporal frames is to more exploit the spacial correlation of the video cube frames extremities. In this way, "Accordion representation" also minimizes the distances between the pixels correlated in the source. That's could be clearer in figure 5:

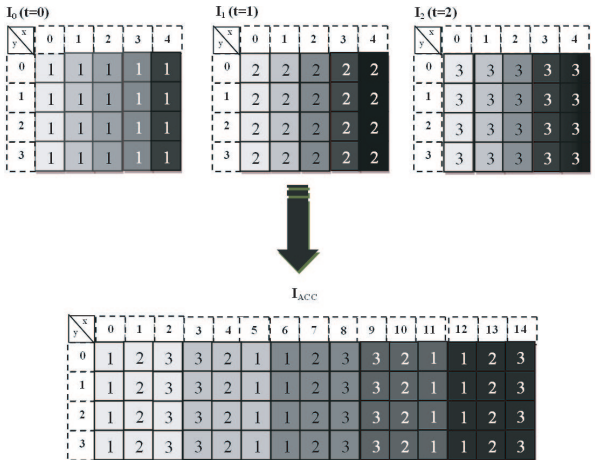

Fig. 5. ACCORDION representation example

Figure 6 shows the strong correlation obtained in the "Accordion representation" made of 4 frames which are extracted from "Miss America" sequence.

Fig. 6. Accordion representation example (Miss America)

C. "Accordion" analytic representation

The "Accordion representation" is obtained following a process having as input the GOP frames(I 1..N) and has as output the resulting frame $IACC$. The inverse process has as input the $IACC$ frame and as output the coded frames (I 1..N). The analysis of these two processes leads to the following algorithms:
The algorithm 1 describes how to make "Accordion representation" (labeled ACC), The algorithm 2 represents the process inverse (labeled ACC^{-1}).

Let us note that:

1) L and H are respectively the length and the height of the video source frames.
2) NR is the number of frames of a GOP.

Algorithm 1 Algorithm of ACC:

```
for xfrom0 to (L * N) − 1 do
  for yfrom0 to H − 1 do
    if ((x div N) mod 2) != 0 then
      n=(N-1) - (x mod N)
    else
      n=x mod N
    end if
    IACC(x, y)= In (x div N,y)
  end for
end for
```

Algorithm 2 Algorithm of ACC $^{-1}$:

```
for nfrom0 to N − 1 do
  for xfrom0 to L − 1 do
    for yfrom0 to H − 1 do
      if (x mod 2) != 0 then
        XACC= (N -1) - n (x*N)
      else
        XACC= n(x * N)
      end if
      In(x,y)=IACC(XACC, y)
    end for
  end for
end for
```

3) $IACC(x, y)$ is the intensity of the pixel which is situated in "$IACC$" frame with the co-ordinates x,y according to "Accordion representation" repair.
4) $In(x, y)$ is the intensity of pixel situated in the Nth frame in original video source.

We can also present the "Accordion Representation" with the following formulas:

ACC formulas:

$$IACC = In(xdivN, y) \tag{2}$$

with n=((x div N) mod2)(N-1) + 1-2((x div N) mod 2)(x mod N)

ACC inverse formulas:

$$In(x, y) = IACC(XACC, y) \tag{3}$$

with XACC = ((x div N) mod 2)(N-1) + n(1-2('x div N) mod 2))+x

In the following, we will present the diagram of coding based on the "Accordion representation".

D. Diagram of coding $ACC - JPEG$

$ACC - JPEG$ Coding is proceeded as follows:

1) Decomposition of the video in groups of frames (GOP).
2) "Accordion Representation" of the GOP.
3) Decomposition of the resulting "$IACC$" frame into 8x8 blocks.
4) For each 8x8 block:
   - Discrete cosine Transformation (DCT).
   - Quantification of the obtained coefficients.
   - Course in Zigzag of the quantized coefficients.
   - Entropic Coding of the coefficients (RLE, Huffman)

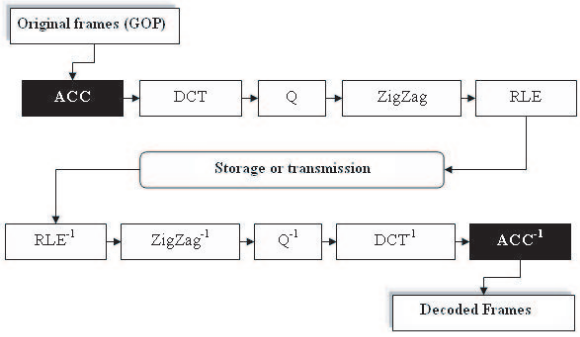

Fig. 7. $ACC - JPEG$ diagram coding

IV. Experiments

Many experiences had been conducted in order to study the performances of our method; we had chosen different kinds of benchmarks. In the following, we summarize the experimental results with some analysis and comments.

A. parameters of the representation

We start by studying the performances of the proposed method with different NR values, it's pointed out that NR presents the number of frames of the video cube that forms the "$IACC$" frame. The best compression rate is obtained with NR=8. Since JPEG process starts with breaking up the image into $8x8$ block, the "Accordion representation" does not have any interest with a GOP made up more than 8 frames. Figure 8 presents $ACC - JPEG\ PSNR$ curves variation according to used NR parameter for "Miss America" sequence(CIF 25Hz).

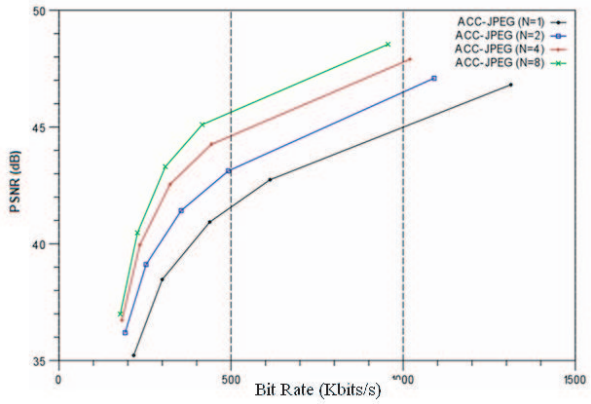

Fig. 8. $ACC - JPEG$ PSNR curve variation according to NR parameter (Miss America)

These results reveals the NR influence on $ACC - JPEG$ compression performance. By multiplying the NR value by

2, the PSNR increases from 1 to 2 dB. This compression improvement is due to the exploitation of the temporal redundancies which become more significant by increasing the GOP's frames number. For NR=1, it acts as the MJPEG which does not exploit the temporal redundancies. By increasing the value of NR, the coder exploits more the temporal redundancies and so offers a better compression performance.

B. Compression performance

In all studied sequences, the $ACC - JPEG$ outperforms the MJPEG in low and high bit rates, it outperforms MJPEG 2000 in high bit rates (from 750 kbs) and it starts reaching the MPEG 4 performance in bit rates higher than 2000 Kb/s. Among the studied sequences, we have got worst compression performance with "Foreman" sequence. The "Foreman" sequence contains more motion than the other studied sequences. This sequence contains non-uniform and fast motion which are caused by the camera as well as the man's face movement. The $ACC - JPEG$ efficiency decreases, measured PSNR is relatively low with an alternate character, especially in low bit rate. In fact, such results are expected as $ACC - JPEG$ eliminate "$IACC$" frame's high frequency data which actually represent the high temporal frequency produced by the fast motion in the Foreman sequence. "Hall monitor" sequence seems to involve less motion compared to the Foreman sequence; the motion takes place only in a very concentrated area. Due to the little amount of motion taking place on the overall image, we observed that our method get better results. Figure 9 shows results of PSNR based comparative study between $ACC - JPEG$ and different existing video compression standards relative to "hall monitor" sequence. The best

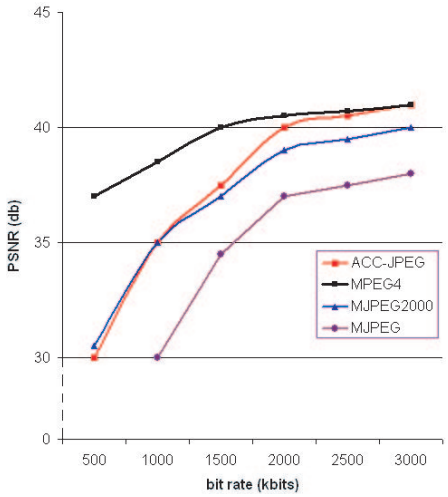

Fig. 9. PSNR evaluation (Hall monitor)

results were given with "Miss America" sequence; "Miss America" is a low motion sequence. The motion is confined to the person's lips and head. Since motion is low, temporal redundancy is high and it is expected that $ACC - JPEG$ becomes efficient.

C. $ACC - JPEG$ artifacts

In the proposed method, the DCT is exploited in temporal domain. Some artifacts produced by 3D DCT based compression methods [9][10] persists in $ACC - JPEG$. Actually, the application of the DCT on $IACC$ allows the transformation from the spatial domain to the frequency domain. After quantification process, we will eliminate the high spatial frequencies of "$IACC$" frame which actually present the high temporal frequencies of the 3D signal source. Thus, a strong quantification will not affect the quality of image but will rather affect the fluidity of the video. The change in the value of a particular pixel from one frame to another can be interpreted as a high frequency in the time domain. Once some of the coefficients have been quantized (set to zero) the signal is smoothed out. Thus some fast changes over time is somewhat distorted which explain the alternate character of the $ACC-JPEG$ PSNR waveform shown in figure 10. However,

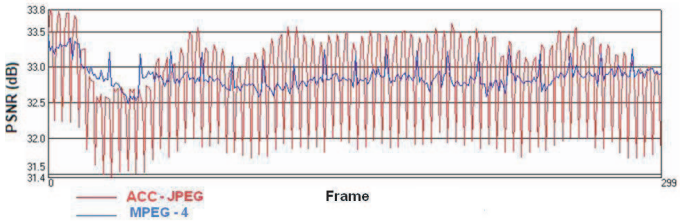

Fig. 10. PSNR waveform comparison between $ACC-JPEG$ and MPEG-4 (Miss America)

some sudden pixels change will be eliminated. This will offer a useful functionality such as the noise removal; Indeed, the very high temporal frequency (sudden change of a pixels value over time) is generally interpreted as a noise. Moreover, some artifacts existing in DCT based compression methods such as spatial distortions generated through the massive elimination of the high spatial frequencies (macroblocking) does not exist in the proposed method as shown in figure 11. The PSNR

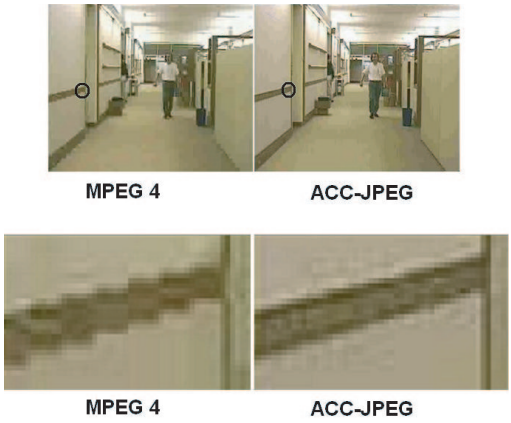

Fig. 11. MPEG-4 Vs $ACC - JPEG$

Curve relative to the $ACC-JPEG$ coding is in continuous alternation from one frame to another with a variation between 31.4 dB and 33.8 dB unlike MPEG PSNR which is almost stable. In one hand, $ACC-JPEG$ affects the quality of some frames of a GOP, but on the other hand, it provides relevant quality frames in the same GOP, while MPEG produces frames practically of the same quality. In video compression, such feature could be useful for video surveillance field; Generally, we just need some good quality frames in a GOP to identify the objects (i. e. person recognition) rather than medium quality for all the frames.

V. $ACC-JPEG$ FEATURE ANALYSIS

The proposed method presents several advantages:

-**Symmetry**: On the contrary of coding schemes based on motion estimation and compensation whose coding is more complex than decoding, the proposed encoder and decoder are symmetric with almost identical structure and complexity, which facilitates their joint implementation.

-**Simplicity**: The proposed method transform the 3D features to 2D ones, which enormously reduce the processing complexity. Moreover, The complexity is independent of the compression ratio and motions.

-**Objectivity**: Unlike 3D methods that treat temporal and spatial redundancies in the same way, the proposed method is rather "selective", it exploits the temporal redundancies more than the space redundancies; what is more objective and more efficient.

-**Flexibility**: The parameters of the $ACC-JPEG$ offer a flexibility that makes it possible to be adapted to different requirements of video applications: The latency time, the compression ratio and the size of required memory depend on the value of the NR parameter. Indeed, by increasing the value of NR, the compression ratio, the latency time and the reserved memory increase. This parameter allows to optimize the Compression/Quality compromise while taking in consideration memory and latency constraints.

-**Random Access:** 3D transform and motion estimation based video compression methods require all the frames of the GOP to allow the random access to different frames. However, the proposed method allows the random frame access. the ACC formula makes it possible to code and/or decode a well defined zone of the GOP (Partial coding).

As conclusion, we can state that the $ACC-JPEG$ is very efficient for scenes with a translatoric character [9], or with slow motion, especially without change of video plan. However, it loses much of its efficiency in scenes with extremely fast moving objects and very fast change of video plan. The ACC-JPEC produces images whose details are clearer and without macro-blocking. According to the particular features of $ACC-JPEG$ quoted in this section, and others in section 4 (Alternate PSNR) it seems that $ACC-JPEG$ can be very suitable to video surveillance applications. In fact, in such applications, we find usually video with uniform motion because used cameras are always fixed on specific supports. With such given video, $ACC-JPEG$ becomes very efficient, it gives good visual quality with clear image details and identifiable moving objects by exploiting the high quality of some frames in the GOP for further recognizing operations. Furthermore, $ACC-JPEG$ seems to be well adapted to embedded or portable video devices such as the IP cameras thanks to its flexibility and its operating simplicity.

VI. CONCLUSION

The video signal has high temporal redundancies between a number of frames and this redundancy has not been exploited enough by current video compression technics. In this research, we suggest a new video compression method which exploits objectively the temporal redundancy. With the apparent gains in compression efficiency we foresee that the proposed method could open new horizons in video compression domain; it strongly exploits temporal redundancy with the minimum of processing complexity which facilitates its implementation in video embedded systems. It presents some useful functions and features which can be exploited in some domains as video surveillance. In high bit rate, it gives the best compromise between quality and complexity. It provides better performance than MJPEG and MJPEG2000 almost in different bit rate values. Over "2000kb/s" bit rate values our compression method performance becomes comparable to the MPEG 4 especially for low motion sequences. There are various directions for future investigations. First of all, we would like to explore others possibilities of video representation. Another direction could be to combine "Accordion representation" with other transformations such as wavelet transformation. The latter allows a global processing on the whole of the "Accordion representation",on the contrary of the DCT which generally acts on blocks.

REFERENCES

[1] E. Q. L. X. Zhou and Y. Chen, "Implementation of h.264 decoder on general purpose processors with media instructions," in *SPIE Conf. on Image and Video Communications and Processing*, (Santa Clara, CA), pp. 224–235, Jan 2003.

[2] M. B. T. Q. N. A. Molino, F. Vacca, "Low complexity video codec for mobile video conferencing," in *Eur. Signal Processing Conf. (EUSIPCO)*, (Vienna, Austria), pp. 665–668, Sept 2004.

[3] S. B. Gokturk and A. M. Aaron, "Applying 3d methods to video for compression," in *Digital Video Processing (EE392J) Projects Winter Quarter*, 2002.

[4] T. Fryza, *Compression of Video Signals by 3D-DCT Transform*. Diploma thesis, Institute of Radio Electronics, FEKT Brno University of Technology, Czech Republic, 2002.

[5] G. M.P. Servais, "Video compression using the three dimensional discrete cosine transform," in *Proc.COMSIG*, pp. 27–32, 1997.

[6] R. A.Burg, "A 3d-dct real-time video compression system for low complexity singlechip vlsi implementation," in *the Mobile Multimedia Conf. (MoMuC)*, 2000.

[7] A. N. N. T. R. K.R., "Discrete cosine transforms," in *IEEE transactions on computing*, pp. 90–93, 1974.

[8] T.Fryza and S.Hanus, "Video signals transparency in consequence of 3d-dct transform," in *Radioelektronika 2003 Conference Proceedings*, (Brno,Czech Republic), pp. 127–130, 2003.

[9] N. Boinovi and J. Konrad, "Motion analysis in 3d dct domain and its application to video coding," vol. 20, pp. 510–528, 2005.

[10] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the dct coefficient distributions for images," vol. 9, pp. 1661–1666, 2000.