Telecommunication Standardization Sector
Study Group 15
Experts Group for ATM Video Coding
(Rapporteur's Group on part of Q.2/15)

Document AVC-575
September 2-3, 1993

Source: PTT research

Title: Video Coding Aspects of the Bellcore 4 QCIF to CIF mixing proposal

Purpose: D,P

In this paper a summary of the comments to the Belcore 4 QCIF to CIF mixing is given (Documents AVC 500 and AVC 501).

The active contributors were:

NTT labs
BT labs
BELCORE
Dornier
SEL Alcatel
PTT Research

Video Coding Aspects of the Bellcore 4 QCIF to CIF mixing proposal

1. Introduction

This discussion is in response to the request of the Tokyo Reporter's meeting. A brief review of the documents gave the following comments;

- Technical feasibility should be examined in the light of overall system configuration. Existing terminals should be accommodated without modification as far as possible.

- Comparison should be made against transcoding MCU in terms of picture quality, delay and other factors as listed in AVC-501.

The meeting decided to continue the study toward the next meeting and the SG15 meeting (both in September) through correspondence.

The task is to draft a reply to WP15/1 with respect to video coding techniques which are necessary for terminals and MCU to constitute the video mixing multipoint system.

2. Questions to be answered

The question we should answer is whether the video coding solution proposed by Bellcore as in AVC-501 (or modified if necessary) can be practical so that SG15 can recommend this method for the continuous presence multipoint system. More specifically,

1) What does this method require for the video coding/decoding in the terminal?

2) What does this method require for the video signal processing in the MCU?

3) What system performance can we obtain in terms of picture quality and delay? Actually unless requantization process is involved, the picture quality of each constituent QCIF picture is maintained even after video mixing (but motion rendition may be affected?) Our prime concern will be buffering delay at the output of the video mixer.

4) What can be gained with this method? This is a more service related question, but can not be separated from the video related questions.

3 Service Aspects

When participating in a conference it is very helpful to be able to see all the participating sites, but in many situations conversation is carried on between two parties with other parties listening in. During the conference the dominant parties may change and at times several parties may contribute, but not all the time. Ideally, each site would be able to see all other sites to have a feeling of their presence, whilst also seeing the current speaker in more detail. So for example the current speaker would appear as a CIF image on one screen and up to 4 locations would appear combined in 1 CIF picture on a second screen. (bt2a)

Most future terminal equipment are in multipoint conferences are likely to be workstation or PC based systems. These will provide dialogue functions

via a "common screen" or a "shared screen" and just seeing the other partners will just be only one aspect of the conference.
Demonstration systems for multipoint video conferences between 4 to 6 partners as shown by different companies (e.g. DEC spin, SUN spin) already show that participants can be presented with sufficient visual quality in relatively small formats as 'background arrangements', leaving room for a common object of interest in the form of: a document, graphic, dialogue cursor/pointer or for the actual speaker.
Future MCU's are required to support:
- the dialogue control
- dialogue (foreground image switching on demand on speaker control)
- Common screen encoding (Dornier)

4. Approaches to the multipoint Systems

There may be two approaches;

1) Existing terminals are kept intact and all the necessary measures are provided by the MCU. The current H.231/243 system is based on this approach. Though there have been defined some enhanced functionality's such as chair control on an optional basis, existing H.320 terminals can participate in multipoint communications without any change.

2) We introduce a new terminal design for this purpose. Hence, only new generation terminals can take part in the continuous presence multipoint system.

There is a significant difference between the two approaches from a service point of view. The Bellcore approach seems to be close to the first one, but the terminal is required to operate in R/4R asymmetric video rates. This functionality is not supported by the existing H.320 terminals.

5. impact on equipment

Although the use of the fill bit requires no changes to H.261 as a standard it does constitute a significant change in the way that existing H.261 terminals use the fill bit. Normally the fill bit is only asserted when buffer under flow is imminent. For this reason and for others discussed later in this document it is not possible to implement this proposal with existing terminals without modification to those terminals. It is likely that some existing terminals and H.261 chip sets will not be capable of using the fill bit to deliberately reduce the active video rate. (bt1)

The proposal requires the introduction of new capability and command codes that must be transmitted and understood by the video terminals. This precludes the use of existing terminals without an update, if possible, to their H.242/H.230 software. As mentioned above the terminals also have to be modified to use the error corrector fill bit for quarter rate video. (BT9)

6. Some observations

6.1 Buffering and hypothetical reference decoder

There is an extreme case where the video mixing as in AVC-501 can work

properly;

- Picture frequencies are locked to each other among the terminals.

- Coded picture rates of different input QCIFs are exactly the same (30Hz, 15Hz, 10Hz, or 7.5Hz) and constant.

- Number of bits per QCIF frame are the same throughout each sequence and among the sequences.

In this case it is obvious that the video bridge can multiplex 4 QCIF coded data into a CIF coded data with minimum buffering delay (one picture time delay may be incurred to accommodate different phases of the input pictures). This case is illustrated in the following diagram with an assumption that all the input QCIFs are synchronised in picture phase. Unit 1 indicates number of bits per QCIF picture. The number of bits of each output CIF picture is exactly 4 units. In this case the video mixing buffer will work with no problems and no delay is involved. HRD specification is also met, thus the existing CIF decoder can decode this split-screen picture.

```
      ---------------------------------------------------------
In-1 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
In-2 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
In-3 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
In-4 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
      ---------------------------------------------------------
Out   | 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4| 4|
      ---------------------------------------------------------
```

A next extreme, but not the most extreme, case is as before but In-4 includes a scene cut, thus intra picture with 12 unit of coded data bits.

```
       ---------------------------------------------------------
In-1 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
In-2 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
In-3 | 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
In-4 | 1|                12                  | 1| 1| 1| 1| 1| 1|
      ---------------------------------------------------------
      ---------------------------------------------------------
Out   | 4| 3| 3| 3| 3| 3| 3| 3| 3| 3| 3| 3|15| 4| 4| 4| 4| 4| 4|
      ---------------------------------------------------------
Out   | 4|3*|3*|3*|3*|3*|3*|3*|3*|3*|3*|3*|15       | 4| 4| 4| 4| 4| 4|

       * stuffing bits
                                                <------->

                                                  delay
```

In this case, the output picture with 15 unit of coded data causes stationary delay of 15/4 picture time. Furthermore, the 4 picture time HRD specification will be violated if the intra picture becomes larger than 16.

There are complicated situations in between as combination of asynchronous nature of input pictures and different and time varying coded picture rates (within a limitation of defined minimum coded picture interval) among input pictures. The stationary video processing delay may vary from almost zero to the worst case mentioned above. Unfortunately there is no time to analyse these intermediate situations.

An essential problem is that the video mixer is passive, hence it has very limited power to control the output buffer occupancy or bit generation

there. Once a large picture arrives, it causes corresponding stationary delay at the output of the video mixer. In case of point-to-point communications, we can isolate the large picture delay by using picture skipping, thus keeping stationary state delay small. (NTT)

6.2 Synchronisation of the data streams

The video data streams have to be synchronised before they can be mixed. In order to synchronise the video data streams the mixer has to wait for new picture start codes in each of the involved datastreams and then has to wait until it has received the complete pictures before the mixed pictures can be redistributed. Or it can send a new picture when it has received a new full picture in one of the channels. In both cases you need large buffers in the video mixer because:

1. The pictures in the different video data streams do not start simultaneously but have offsets to each other.

2. The length of the pictures in the different video data streams may vary due to the fact that

a) the contents of the pictures are different, this results in different compression ratio's

b) one picture may be coded in INTER mode, whereas others are coded in INTRA mode,

3) All video pictures which contain INTER-coded blocks can not be left away because they are necessary for the reconstruction at the receiving side.

A sophisticated management mechanism is needed, and requires large buffers.

This leads to a large delay of the video data.

In any case the rate if the CIF output bit rate has to be higher than the sum of the four incoming bit rates to compensate for the stuffing. (PTT)

6.3. More delay considerations

The simulation studies paper demonstrates end to end delays for a 256 kbit/s video bit-rate (64 kbit/s per QCIF source) of between 266.67 and 644.44 ms. The mean delay appears to be approximately 440 ms. At 256 kbit/s the end to end delay for a single CIF encode and decode is approximately 260 ms and does not vary much with picture content. This gives a back to back delay of 520 ms without any optimisation. The proposed method of combination therefore appears to give only a marginal improvement in end to end delay and in some cases causes a degradation. (BT 4)

6.4 Effect of the distribution of FUR's

It is not clear whether the simulations performed (by Belcore) have taken into account the dropping of frames by the source encoders. At low bit rates it is quite normal for an H.261 encoder to drop frames rather than quantise a picture more heavily in order to limit the coded bit-rate. The dropping of frames results in those frames that are transmitted being allocated more bits than average. The effect of this will be similar to the Intra coded frame in response to a fast-update request (FUR). Unlike the FUR the timing of the dropped and transmitted frames can not be controlled. It is quite likely that all 4 QCIF sources will drop frames at the same time and then start to code pictures with a large number of bits in phase. This has been shown to result in longer end to end delays. It is questionable whether there would be much gain in distributing the

FURs in this case. (BT 6)

The dropping of frames also has an implication on the way the 4 QCIF pictures are combined. The Bellcore proposal suggests that the Temporal References from all but 1 source are ignored and that pictures are combined as they arrive to achieve minimal delay. With dropped frames this could result in temporal distortion. (BT7)

6.5 Synchronisation of terminal Clocks

Document AVC-501. section 4.2 states:

"The videoconferencing terminal clocks can be locked to the network clock so that data transmission is synchronous between the terminals and the network. These synchronous terminal clocks also ensure the frame rates of different QCIFs are exactly the same but their frame phases may be different."

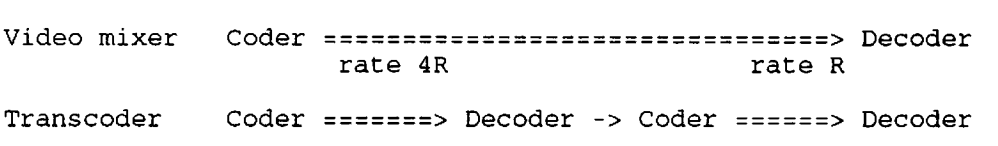
Such locking is not common-place in terminal designs. BT is not aware of any which incorporate it. While it is technically feasible, it does place restrictions on the selection or design of cameras. For integrated terminals these may be under the control of the designer but for additional plug in cameras and other sources such as video tape playback machines, the locking requirement is a drawback. (BT 8)

7. Picture quality

The simulation study paper (AVC-501) concedes that although the pel-domain combining results in slightly degraded picture quality, the degradation is insignificant and almost undetectable subjectively. This conclusion is based on comparing the picture quality before and after a decoded picture has been re-coded and decoded a second time. A more relevant comparison might be between the following two cases. The first is a double encode and decode entailing an initial encoding and decoding using CIF at the full bit rate available from a terminal followed by coding and decoding with QCIF at one quarter of the bit rate. The second case is the single encode and decode with QCIF and quarter bit rate which corresponds to the Bellcore proposal. It is expected that the objective and subjective degradation's introduced by the double encoding will be significantly less than those observed in the Bellcore experiments. (BT 3)

One of the features of the video mixer is that it does not involve picture quality degradation at MCU. Instead of using rate R toward the MCU, we can use rate 4R for the transcoding MCU. This may reduce picture quality degradation due to transcoding It is quite obvious that if the rate of the channel toward MCU is very high, there is no degradation involved at all.

The comparison should be between the following;



the chosen rate should range from 19.6 kbit/s up to 480 kbit/s (ISDN rates 2B, 6B and primary access)
The transcoding MCU presumably places no limitations to the operating mode

of the terminal. If the picture quality and delay are comparable, this can be a practical solution as well for continuous presence multipoint Systems. (NTT)

With the Bellcore proposal only a low bit rate QCIF picture is available from each terminal. For a 2B conference the video rate, assuming 48 kbit/s G.722 audio, no data and no encryption, is 78.4 kbit/s. Each individual QCIF rate will be limited to 19.6 kbit/s. The display of this picture as a full size image on anything other than a small screen would not be acceptable. (This can be confirmed by pictures seen at the meetings of the Rapporteur for Very Low Bit Rate Coding.) (BT 2b)

Experiments carried out SEL Alcatel using this method on 1B (4R = 64 kbit/s ???), lead to unacceptable results, because the four QCIF images gave an image quality which was too poor for the multipoint service.(SEL)

8 ALTERNATIVE METHODES

The hardware to support the combination of pictures at the GOB level may not be as complex as that required to decode the picture and re-code it but all of the complexity saving lies in the compression process for which chip sets and complete codecs are readily available. (Line terminating, clock extraction, data extraction etc. are necessary for both approaches.) The cost of an H.261 encoder and decoder is decreasing all the time, and soon its marginal cost over finding and manipulating the picture and GOB headers and controlling the input buffers will be insignificant in the overall consideration. It is arguably a more complex task to build a 4 QCIF to CIF combiner than to put a decoders back to back with an encoder through a 4:1 spatial decimator/combiner. (BT 5)

9. Impact of the method

1) What does this method require for the video coding/decoding in the terminal?
The video part of the terminal (H.261) has to be changed to asymmetric mode.
When delay needs to be reduced definition of Hypothetical Reference Decoder (HRD) needs to be changed.

2) What does this method require for the video signal processing in the MCU?
The proposed method avoids decoding, recoding. The mixing can be carried out in the H.261 domain and is relatively simple.
However todays technology propides single chip solutions for H.261 encoders and decoders.

3) What system performance can we obtain in terms of picture quality and delay?

Delay:
If the definition of hypothetical reference decoder is not changed, the method proposed by Belcore gives a allay which is the same range as transcoding (or might in some cases even be higher)

Picture quality:
Although intrinsically the proposed method gives no degradation, the limitation to QCIF and a bit rate of R in stead of 4R.
A fair comparison has not been shown yet.
Experiments on 1B (total bit rate 64 kbit/s) carried out by SEL Alcatel gave an unacceptable picture quality (frame rate droped below 1 frame/s).

A 2B operation (128 kbit/s using G.722 for audio)would give an unacceptable quality (extrapolation from results shown in the meetings of the Rapporteur for Very Low Bit Rate Coding)

4) What can be gained with this method?
Obviously continues presence is a highly desirable feature in multipoint conferencing. For the service however other advanced display methods such as e.g.

- a common screen (in high quality), or
- a high quality picture of the person who is speaking combined with a lower quality picture of the other participants

may be preferable. The proposed method gives not the flexibility required for these display method. A transcoding method can provide these features with no limitations.

10. Conclusion

The method proposed by Belcore is simple in it's basics. It requires small modification to the definition of the terminal. The MCU can do a relatively simple mixing in the H.261 domain. The method gives no degradation due to transcoding.
Drawbacks of the method are:

- Changes to the H.261 terminal equipment that already exists is a problem.
- Technology developments makes transcoding and pel domain mixing cheap.
- The method still gives substantial delay.
- The method gives no flexibility in bit rates and to the users.