ITU - Telecommunications Standardization Sector Document Q15-D-47
STUDY GROUP 16 Filename: q15d47.doc
Video Coding Experts Group (Question 15) Generated: 15 April '98

Fourth Meeting: Tampere, Finland, 21-24 April, 1998

Question: Q.15/SG16

Source: Gunnar Hellström
National Post and Telecom Agency Tel: +46 751 100 501
Box 5398 Fax: +46 8 556 002 06
S-102 49 Stockholm, Sweden Email: gunnar.hellstrom@omnitor.se

Title: Sign language and lip reading application profile.

Purpose: Draft

Draft Application Profile
Sign language and Lip-reading real time conversation application of low bitrate video communication.

Summary

Sign language and lipreading are important application areas for video communication. For successful transmission of the visual language components, certain quality requirements must be met.
This is an application profile document that gives background to the requirements and gives guidance on how they can be met. It does not intend to propose new video coding schemes, but indicate instead how current and planned video coding can be applied for good results in this application area.

Introduction

Millions of deaf people have a Sign language as their first language and are eager to get an opportunity to use it for distant conversations. The conversation speed when using sign language is comparable to the speed in voice conversation.
Persons with hearing loss may get very good clues to speech perception by viewing the face of the speaker for lip-reading.

This document describes factors of importance when applying low bitrate video coding for acceptable performance in sign language and lip-reading.

Contents

1 Scope:

This application profile for sign language and lip-reading gives the characteristics needed from a videocommunication system for person to person conversation in sign language and in lip-reading supported speech.
It sets performance requirements that are important to meet to enable the conversation.
It describes how sign language and lipreading performance can be evaluated.
It also suggests factors to be handled externally to the video coding protocol in the terminal design and even in the environment where terminals are used for sign language and lip-reading.

2 Abbreviations

fps	Frames Per Second (=pictures per second)
CIF	
QCIF	
SQCIF	

3 Definitions

4 Non normative references

[Hellström Delvert Revelius 1996] Quality requirements on Videotelephony for Sign Language
[Sg12] Quality requirements on conversational use of video communication
[Frowein 1993] Videotelephony for lip-reading support
[…..1997] Reference scene for quality evaluation of video communication for sign language and lip reading.

5 Basic needs for reproduction of sign language and lip-reading

5.1 Basic characteristics

The language carrying components in sign language are movements and positions of the hands, the eyes, the mouth, the face and the body.
In lip-reading, it is movements of the face. Often lip-reading is supported by voice. In other cases it is used together with sign language.
In video-coding terms, the scene with one signer or speaker is regarded to contain a medium amount of motion occationally being high.

5.2 Temporal resolution requirements

Both sign language and lip reading need good reproduction of movements.

- Usability for sign language and lip reading is reported to start at 12 pictures per second (fps), with improvements at least up to 20 fps.[Hellström, Delvert, Revelius 1996], [Frowein 1993]

An example of the requirements can be found from fingerspelling, where each letter has a unique hand position and spelling is done by showing thesepositions in rapid sequencce. Fingerspelling is used at rates up to 10 letters per second. For reliable reproduction, at least two pictures per letter should be sampled. That sets the requirement at 20 frames per second.
Signing just occasionally use fingerspelling. The greater part of signing is done with signs for complete concepts (like words). Also during such signing, rapid hand movements occur and short blinks carrying grammatical information. In many cases the temporal resolution requirements are the same as when fingerspelling.

A raw requirement figure for lip-reading can be calculated from phoneme rate in normal speech. A normal phoneme rate is 10 per second. In order for all phonemes to have a visual representation, at least 20 pictures per second should be reproduced.

In both cases, the speed of language production can be slightly reduced by will, and that explains why 12-15 frames per second is found possible to use.

An analysis from the test sequence "Irene", explains this further.

Analysis of the frame rate requirement.

Fingerspelling

This is an approximate representation of two fingerspelling sequences in "Irene". The numbers are frame numbers from the beginning of the MPEG version. The letters indicate when the letters are quite clearly formed by the hand. A dash indicates that no clear letter is formed in transition between letters. The first is "Pia Wickman", with the last "a" only visible on the mouth.

TABLE 1 - Example of fingerspelling representation in frames at 25 frames per second.

"Pia Wickman"		47	-	67	-	"Edsviken"		324	-
		48	-	68	m			325	k
Frame	Letter	49	-	69	-	Frame	Letter	326	k
29	p	50	-	70	n	308	e	327	k
30	p	51	w	71	n	309	e	328	-
31	p	52	w	72	n	310	-	329	e
32	p	53	-	73	n	311	-	330	n
33	-	54	i	74	n	312	d	331	n
34	-	55	-	75	n	313	s	332	n
35	i	56	c	76	n	314	s	333	n
36	i	57	c	77	n	315	s	334	n
37	i	58	c			316	s	335	n
38	i	59	-	. . .		317	-	336	n
39	-	60	-			318	v		
40	a	61	k			319	v		
41	a	62	k			320	v		
42	a	63	k			321	-		
43	a	64	-			322	i		
44	a	65	-			323	-		
45	a	66	-						
46	a								

Among these 18 letters, 5 are clear only on one frame and would risk to be lost at the framerate 12.5 frames per second that would appear if every second frame was skipped in the transmission scheme.

The distribution is:
1 frame - 5 letters
2 frames - 2 letters
3 frames- 4 letters
4 frames- 3 letters

7 frames- 2 letters (ending of phrases)

Mean length inside phrases: 2.3 frames per letter.

There is a grammatical blink in frames 394 and 395 used as a sentence marker.

Conclusion: In this example, the letters within words vary between 1 and 4 frames in time, the frames representing 40 ms each. The mean length is 2.3 frames visibility per letter. The example is too small for making any real statistical conclusion on, but it can be seen that with this fingerspelling speed, a framerate of 12 would require some guesswork to perceive what is fingerspelled.

General signing
Large parts of the clip is signed with signs without fingerspelling, comparable to words.
A simple analysis has been performed on one phrase. The prase is presented here, transcribed sign by sign with the number of frames each sign occupies in paranthesis
The sequence is found between frame 406 and 529 in the MPEG version.

"SHE(7) TELLS(7) HERSELF(11) HOW(4) SHE(2) FELT(11) EXPERIENCED(13) ADOLESCENCE (16)".

Each sign is a motion, so it is definitely required to reproduce it with a framerate enough to bind the result into moving pictures. (15 fps). The conclusion is that the framerate requirements seem to be slightly less than for fingerspelling. No sign in this sequence was shorter than 2 frames and they did not contain more rapid motion than the fingerspelling.

5.3 Spatial resolution requirements

For spatial resolution it is reported that for person-to-person calls, the following is needed: [Kamata], [Frowein], [Hellström,Delvert]

- QCIF is possible to use, but the smallest details, telling about gaze directions are lost, causing stress for the recipient.
- CIF is good and the increase from QCIF to CIF is felt to give better language perception.
- SQCIF is too coarse for reliable perception.
- If different resolution is used in different parts of the picture, it is the hands and the face that needs highest resolution.

A simple theoretical verification can be done. In the head to stomach view usually used in person to person conversation, a finger is approximately 1/50 of the picture width. In order to resolve fingers reliably in a picture, a finger should be represented by at least 3 pixels. That puts the spatial resolution requirement to QCIF, that contains 176 pixels in width. Eye gaze direction is also important and require higher resolution. Therefore CIF is appreciated, and preferred.

5.4 Fidelity

In video communication, blur is often introduced during motion.
VHS video is reported to be sufficient for good perception of sign language and lip reading. In video recordings, rapidly moving objects are often shown with considerable blur because shutter speeds are normally 1/50 to 1/60 of a second. This indicates that blur is acceptable on rapidly moving objects involved in big movements.

Since SQCIF is found too coarse for reliable perception, but still have some readability, the occasionally introduced blur should not go beyond SQCIF in the sign language and lip-reading application.

5.5 Delay

End to end video delay from the sending camera to the receiving display is critical in the conversation application.
Values below 0.4 seconds are preferred . [Sg12]
Values over 0.8 seconds are felt hindering a good conversation. [Hellström Delvert Revelius 1996].

5.6 Synchronism

For hearing supported lip-reading, the synchronism between sound and video is essential. Time differences of up to 100 ms is reported to be acceptable.

5.7 Conclusion

Aim at 25-30 frames per second at CIF resolution and max. 0.4 second delay, accepting a blur less than corresponding to QCIF during medium motion.

Accept, if needed in very low bitrate environments, 12-15 fps QCIF with medium motion and occasional degradation to corresponding to SQCIF during heavy sign language motion.

Keep sound synchronism better than 100 ms.

Make sure that end-to-end delay is kept below 0.4 seconds.

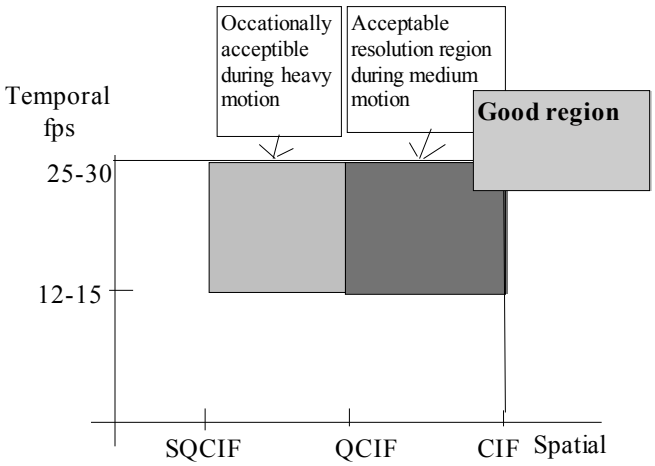

Figure: Resolution requirements for sign language and lip-reading.

6 Performance verification

6.1 Reference material

A sign language clip can be used for evaluation. The clip "Irene" contain suitable amount and rapidity of motion.

6.2 Performance evaluations

A codec or a terminal setup is tested by transmitting the reference scene through a coder or through a set of video phones with a network connection. The result is recorded and evaluated.

The framerate during signing is evaluated.
The selected static resolution is noted.
Any introduced extra blur during medium motion is measured by comparing the recorded frames with pictures from the same scene with resolution reduced to QCIF resp. SQCIF. Blur is only judged on hands and face.

The delay is measured according to the guidelines.
The synchronism is measured.

From these recordings the performance can be evaluated and compared to the goals described above.

7 Advice to the terminal implementor.

- In order to satisfy the users, the terminal should implement a few features.
- It should provide a connector for external alerting systems.
- The users may occasionally need to revert to text conversation. Therefore it should be possible to invoke the text conversations protocol T.134-T.140

8 Advice to the user.

The user should arrange a user environment with good lighting and an even surface behind the camera.