Impact of Adaptation Dimensions on Video Quality

Jens Brandt and Lars Wolf
Institute of Operating Systems and Computer Networks (IBR),
Technische Universität Braunschweig, Germany
{brandt|wolf}@ibr.cs.tu-bs.de

***Abstract*—The number and types of mobile devices which are capable of presenting digital video streams is increasing constantly. In most cases the devices are trade-offs between powerful all-purpose computers and small mobile devices which are ubiquitously available and range from cellular phones to notebooks. This great heterogeneity of mobile devices makes video streaming to such devices a challenging task for content providers. Each single device has its own capabilities and individual requirements, which need to be considered when sending a video stream to it. Thus, to support a great range of different devices, the video streams need to be adapted to the requirements of each device. To get an idea of how different adaptation methods may affect the experience of users watching a streamed video on a mobile device, we inspect the influence of three major adaptation dimensions on the produced quality of the stream. Based on these results, we are able to give a clear recommendation for a multidimensional adaptation of digital video streams.**

I. Introduction

The prominence of digital video on the Internet is rising constantly. Faster Internet connections and more powerful devices make video streaming over the Internet more and more popular. The usability of digital videos that are streamed over a wireless network connection is still limited by the available resources at the receiving device. The presentation, for instance, of a video stream with 1920×1080 pixels on a screen with a resolution of only 320×240 pixels is not possible without a high amount of processing power at the receiving device. The transmission of such a high quality video stream over a wireless link may also consume a large part of the available network capacity, which then may interfere active network connections of other devices. A limited network capacity may additionally prevent the device from decoding and displaying the video stream at all. One solution of these problems is to perform an adaptation of the video stream to the requirements of mobile devices. For this purpose, we already presented an architecture for multidimensional video transcoding [1] that is able to adapt MPEG-4 video streams in the temporal, the spatial as well as in the detail dimension.

Mobile devices typically have small screen sizes, which results in lesser details visible on the screen of mobile devices compared to home entertainment displays. A user of a mobile device may watch a video stream in a variety of situations: while being at home, sitting in a park or, to name but a few, while traveling on a train. Additionally, the users of such devices may vary their viewing angel and distance more often and dynamically, compared, for instance, to watching a video stream on a fixed screen. As a consequence, also the environment of mobile devices such as the audio-visual ambience as well as the network conditions, may vary more often over time compared to static devices. Thus, the experience of watching video streams on mobile devices differs to a great extent from those scenarios related to static displays usually found in the area of home entertainment. For sequences from soccer games McCarthy et al. observed that potential users preferred lower frame rates but higher detail resolution on mobile devices [2]. For other genres similar findings were presented for scalable video coding by Eichhorn and Ni in [3]. Besides concentration solely on the temporal resolution of a video stream, in our work we additionally investigated the effect of adapting the spatial and the detail resolution as well.

II. Video Adaptation for Mobile Devices

The main goal of video adaptation is to produce a video stream which fits to the requirements of the requesting client. In video streaming scenarios, a requested video stream firstly needs to be transmitted to the client over the network. If the bit rate of a video stream is higher than supported by the network connection, the receiving client will not be able to receive the stream properly. In such a situation, the bit rate at which the client can receive the stream from the network is the most limiting requirement. The bit rate of a video stream mainly depends on three different dimensions: the spatial resolution, the temporal resolution and the detail resolution of the stream. However, a certain target bit rate can be achieved by several combinations of adaptation dimensions. The temporal resolution of a video stream, for instance, might be reduced while keeping the detail quality of the remaining frames. Another possibility would be to reduce the detail quality while keeping the frame rate of the stream. Both approaches may achieve a similar bit rate reduction, and it needs to be identified which approach produces a better quality. For the spatial resolution, a similar situation exists. A video stream might be downscaled to the resolution of the receiving device or even further to retain a higher detail quality for each single frame. Altogether there are $2^3 = 8$ different combinations of the three mentioned adaptation dimensions which may achieve a lower bit rate. Additionally, for each dimension different adaptation parameters may exist. Each of these dimensions and parameters might lead to different other limitations and will be discussed in the next sections.

If the capacity of the network connection is not the limiting factor, for instance in a home entertainment scenario with a network connection that supports high bit rates between the video source and the client, the situation is much easier as the temporal and the spatial resolution of a video stream can be adapted independently. If the display resolution of the requesting client is the limiting factor, the spatial resolution of the video stream needs to be tailored accordingly. If the temporal resolution of the stream is the only or an additional limiting factor, the frame rate needs to be reduced. In both situations, the detail quality is not reduced as the available bit rate of the network connection is not the limiting factor.

A. Spatial Adaptation

Adaptation of the spatial resolution can be used to reduce the bit rate of the stream and to meet the resolution of the client display. The latter aspect can be optimally achieved by reducing the spatial resolution of the stream to exactly the display resolution of the receiving device. However, better quality results might be achieved by reducing the spatial resolution further while keeping the detail resolution at a higher level. Another possibility is to keep the spatial resolution higher than the display resolution while reducing the detail resolution of each frame. Thus, three different possibilities for the target resolution can be distinguished:

i) The target resolution is higher than the display resolution.
ii) The target resolution is the same as the display resolution.
iii) The target resolution is lower than the display resolution.

To identify which target resolution produces the best quality, we encoded ten well-known video sequences (i.e., akiyo, deadline, mobile, etc.) at different spatial resolutions and bit rates. For each encoded sequence, we evaluated the produced quality in terms of the average Y-PSNR values with respect to two different target resolutions: i) CIF resolution with 352×288 pixels and ii) a resolution of 264×216 pixels, which is CIF downscaled by a factor of 0.75 on both axes.

Each video sequence was encoded at different bit rates ranging from 40 kbit/s to 480 kbit/s with a temporal resolution of 25 frames per second and three different spatial resolutions, i.e., CIF resolution, a resolution of 264×216 pixels and QCIF resolution at 176×144 pixels. For the encoding process we used the MEncoder from the MPlayer project[1] in combination with the MPEG-4 codec from the FFmpeg project[2]. The average Y-PSNR values were calculated with respect to the target resolution. Therefore, the MPlayer was used for decoding the video frames and the PSNR values of each decoded frame were computed by the use of some tools from the Netpbm project[3]. For the target resolution of 352×288 pixels, i.e., for the CIF resolution, we calculated the PSNR values from the decoded and upscaled pictures. For the second target resolution of 264×216 pixels, we firstly downscaled the version encoded at CIF resolution to fit to 264×216 pixels, which simulates the scaling process necessary on the decoding device with the given target resolution. Afterwards, all versions were upscaled to CIF resolution again in order to calculate the PSNR values.

Figure 1 illustrates the processes used to create the different versions of the stream as well as the versions used to compute the PSNR values. For the target resolution of 264×216 pixels the lower path for producing the stream at CIF resolution is used whereas for the target resolution of 352×288 pixels the upper path is used.

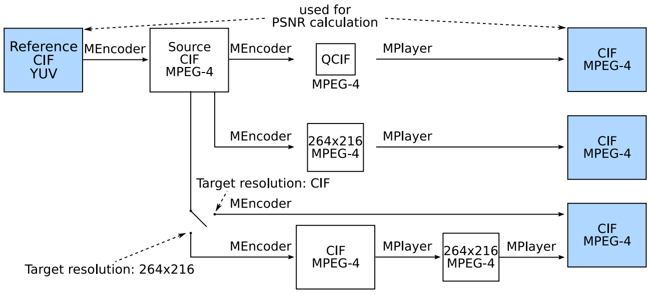

Fig. 1. Evaluation process for different spatial target resolutions

For each video sequence there exists a lower and an upper bound for the bit rate that can be achieved by the used encoder at the given spatial and temporal resolution. At the lower bound the encoder uses the highest possible quantizer scale value and therefore produces the lowest possible quality. At the upper bound the encoder accordingly uses the lowest possible quantizer scale value and produces the highest possible quality. The values of these bounds depend on the characteristics of each video sequence and therefore, the following graphs do not always contain PSNR values for the full range of bit rates between 40 and 480 kbit/s.

Figure 2 exemplarily shows the average Y-PSNR values of the deadline sequence for each target resolution compared to the bit rate of each stream. It can be observed that in both situations a lower resolution than the target resolution results in substantial lower PSNR values for all bit rates. For the lower target resolution of 264×216 pixels, it can be further seen that encoding at a higher resolution slightly increases the produced quality. The reason for this increase is that the motion estimation in the encoder benefits from the higher resolution. However, this quality increase resulting from the

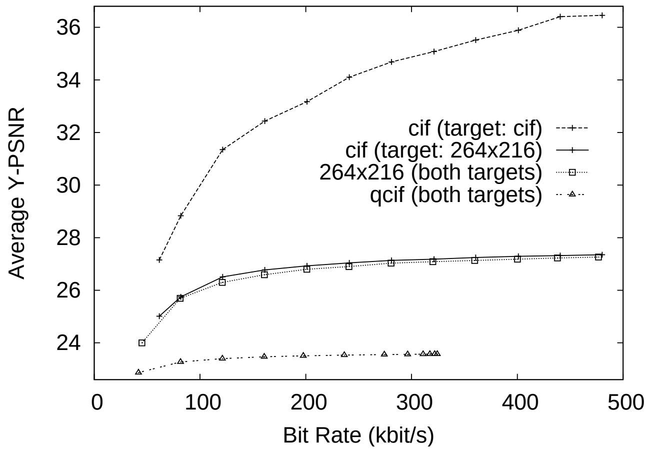

Fig. 2. Video quality at different spatial resolutions - deadline sequence

[1] http://www.mplayerhq.hu
[2] http://www.ffmpeg.org
[3] http://netpbm.sourceforge.net

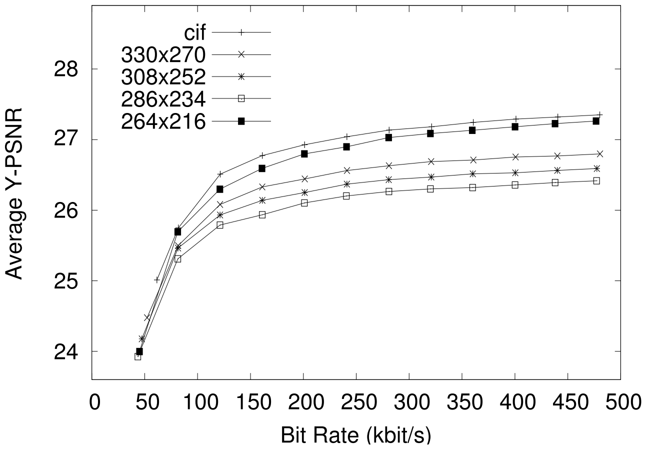

Fig. 3. Video quality at different spatial resolutions - deadline sequence

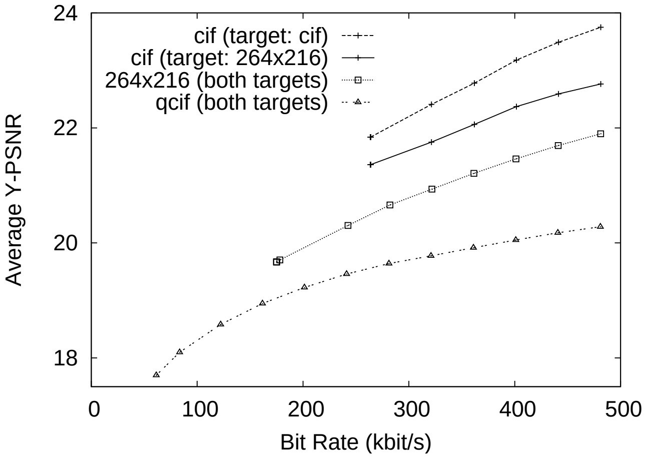

Fig. 4. Video quality at different spatial resolutions - mobile sequence

higher resolution is not significant. To further evaluate the better quality of the stream encoded at a higher resolution we produced additional video streams at intermediate resolutions between 264×216 and 352×288 pixels. Figure 3 shows the PSNR values for these intermediate resolutions. For bit rates above 40 kbit/s it can be clearly seen that the version of the stream with the same resolution as the target resolution of 264×216 pixels produces better results compared to those versions with higher resolutions, except the original resolution of 352×288 pixels.

The results for most of the other sequences show very similar results to those of the deadline sequence. Only for some sequences we observed that the PSNR values of the version with a resolution higher than the target resolution were about 1 dB increased than the PSNR values of the version at the target resolution. As both the sequence with the lowest amount of motion and visible details as well as the sequence with the highest amount of motion and visible details showed this effect of higher PSNR values, we could not clearly identify any similarities in the affected sequences. Figure 4 shows the results of the mobile sequence as an example of such a sequence.

A screen resolution of 352×288 pixels is quite low and may not be representative for the great range of different devices. Therefore, we also evaluated the produced quality for a higher resolution of 704×576 pixels. This resolution is four times higher than CIF resolution and is therefore also called 4CIF. For this resolution we used two further test sequences as the previously used sequences were only available at a maximum resolution of CIF. One is called the *harbour* sequence and shows some slowly moving boats in a small harbor. Due to the great amount of moving objects in this sequence it has similar characteristics as the mobile sequence. The second sequence is called *soccer* and shows some soccer players on the playing field. From the amount of motion it is comparable to the foreman sequence. For both sequences, we evaluated the produced quality at a target resolution of 528×432 pixels which is the original resolution downscaled by a factor of 0.75. In this case it was sufficient to concentrate on only one target resolution that is lower than the original resolution of the streams because the results for the target resolution of 352x288 pixels showed very clearly that reducing the spatial resolution any further than needed for the target resolution produces bad quality results.

The results for both sequences are very similar to the previous ones: The differences between the PSNR values are quite small for the versions at the original and at the target resolution. The only difference that we found is that for low bit rates (i.e., below 2500 kbit/s) the PSNR values of the video stream with the target resolution are between 0.1 and 0.5 dB above those of the stream with the original resolution. This may result from the greater range of bit rates that can be achieved for the higher resolution sequences. For the soccer sequences we achieved bit rates up to 14000 kbit/s and for the harbour sequence we achieved even higher rates.

However, as mentioned before, there are other resource limitations which need to be considered as well. When using another spatial resolution than the target resolution, more processor cycles are needed for the spatial scaling at the receiving client, which usually results in a higher energy consumption at the device. This could only be justified by a substantial quality increase. Our evaluation, however, did not show a significant increase of the produced quality in the cases where the video was encoded at a resolution different from the target resolution.

In summary, it can be observed that any reduction of the spatial resolution other than that needed to adapt a video stream to the display resolution of the receiving device is usually not sensible with respect to the quality of the stream. With respect to the bit rate of the video stream, a reduction of the resolution more than needed to meet the display resolution may be necessary in order to reduce the bit rate of the stream as needed. However, this might result in additional quality loss and further processing cycles needed for any spatial scaling process on the receiving device.

B. Temporal Adaptation

Another possibility to reduce the bit rate of a video stream in order to meet the requirements of a requesting client, is the reduction of the temporal resolution. In contrast to an adaptation of the spatial resolution, there typically is no temporal target resolution which needs to be achieved. Unless the requesting device has any frame rate limit, the only criteria which can be used to identify a reasonable frame rate is the produced quality. Therefore, we need to determine if the quality of a stream can be increased by reducing the frame rate rather than reducing the detail resolution, which usually is used to reduce the bit rate of the stream.

In the context of temporal adaptation, the analysis of PSNR values would not help to evaluate and compare the produced quality, because the PSNR values would be computed from streams at different frame rates. PSNR values are computed frame-by-frame with respect to a reference stream. If the produced stream has a frame rate lower than the reference stream, the missing frames need to be interpolated. This interpolation usually results in poor PSNR values if there is some amount of motion in the sequence. A second possibility is to compute the PSNR values for the resulting frames only. This method usually results in higher PSNR values if the frame rate is reduced. Therefore, we conducted interviews with potential users and evaluated their responses to different test videos. We encoded different sequences at different frame rates and presented them to potential users on a mobile device with a 3.5 inch display and a resolution of 320×240 pixels.

We chose four different video sequences which cover a broad range of different characteristics such as genre, the amount of motion, the number of scene cuts and regions of interest. The first sequence was taken from a news broadcast, showing a speaker in front of a static background intercepted by short news clips. This is a typical news sequence including parts with a low amount of motion alternating with passages with higher amount of motion. The second sequence is a short section from a soccer game broadcasted on TV. In the beginning of this sequence, the playing field is shown from the perspective of the audience followed by some close-ups of singe players and groups of players. The characteristics of this sequence are quite different from the news sequence as there is a high amount of movement both in the foreground and in the background. This sequence is also a good example for a small region of interest as most people who are watching a soccer game usually are interested in movements of the ball. The third and fourth test sequences were chosen from movie trailers available on the Internet: one sequence was taken from an animation movie and another one was taken from a movie with natural video content. Both sequences have a high number of cuts as well as a high amount of motion but differ in the genre. The length of all four sequences was between 75 and 90 seconds. All sequences were encoded with a fixed spatial resolution which fitted best to the screen of the mobile device used in the test situations and a constant bit rate of 180 kbit/s, which is a typical rate for videos on the Internet [4].

As the temporal resolution, we used three different frame rates: the original, 12 or 5 frames per second (fps). The original frame rates for the news and soccer sequences were 25 fps and 24 fps for the movie trailers. All frame rate variants of each sequence were presented to a test person in changing order to avoid any effects resulting from the order of the variants. The video sequences were displayed in full-screen mode on the mobile device that the participants held in their own hands. The environment during the tests included typical indoor and outdoor situations where people may watch videos on a mobile device. Each participant was asked to choose his or her preferred version of the video. Finally, they were also asked to rate the preferred version with a grade from 1 (best) to 6 (worst). The tests were conducted with a total number of 50 non-expert users with ages between 21 and 59 years. Each video sequence was rated by 36 to 50 of these different test persons.

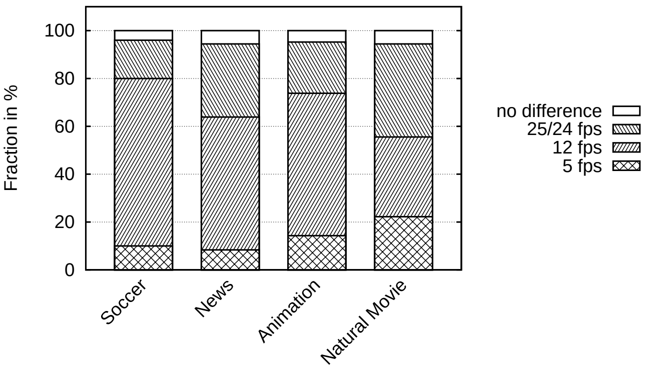

Fig. 5. Frame rates preferred in the video tests

Figure 5 shows the preferred frame rates for each test sequence. For the soccer sequence, 70.0 % of all participants preferred the version with a frame rate of 12 fps, 16.0 % preferred the version with 25 fps, 10.0 % preferred the version with only 5 fps, and 4.0 % of all participants did not notice any differences between the three different versions. For the news sequence as well as for the sequence from the animation movie, we can see very similar results. For the news sequence the version with a frame rate of 12 fps was preferred by 55.56 % of the users and for the animation sequence this version was preferred by 59.52 % of all participants. Thus, for three of the four sequences the version with the halved frame rate of 12 fps was preferred by the majority of the users. Especially in the case of the soccer sequence, a very clear preference of the version with 12 fps can be seen. Only in the case of the natural movie trailer the test persons slightly preferred the version with the higher, i.e., the original frame rate. 38.88 % of the users preferred 25 fps and 33.33 % preferred a rate of 12 fps. Also the amount of users which preferred the version with only 5 fps is noteworthy with 22.22 %. The fact that there is no significantly preferred version of the natural movie sequence may result from the characteristics of this sequence. It has a very high number of scene cuts and fades combined with a high amount of motion. Due to this characteristics, the users might not have noticed all the details between the scene cuts and therefore may prefer the higher frame rate because of a smoother motion within the sequence.

At first glance the results observed for the first three sequences that the users preferred a lower frame rate also in the case of high amount of motion such as in the soccer sequence, seems to be a bit surprising. However, this phenomenon can be explained by the number of details visible in each single frame. If the frame rate of a video stream is reduced while at the same time the bandwidth of the stream is kept, each single frame may consume more bandwidth. If there is more bandwidth available for each frame, the quantizer scale value can be reduced and the frame will contain more visible details. Especially, in video sequences with small moving objects such as a football or the players in a soccer sequence, the visual quality of each single frame seems to be more important than a smooth playback. Similar results for sequences from soccer games as one characteristic type of video sequences were also observed by McCarthy et al. in [2]. For other genres similar findings were presented for scalable video coding by Eichhorn and Ni in [3]. Only in the case when details in a video sequence are not visible any more due to a great amount of motion or a very high number of scene cuts, more test persons prefer a higher frame rate.

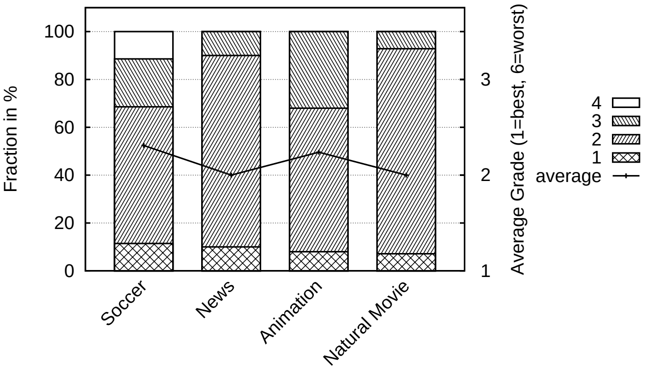

Fig. 6. Grade of preferred video

The users gave their preferred video sequence a grade between 1 (best) and 6 (worst). These grades are shown in the histograms for each video in figure 6. For all four sequences, a dominance of a good grade around 2 can be seen. The average grades given for the preferred version of the sequences are 2.31 for the soccer sequence, 2.0 for the news sequence, 2.24 for the sequence from the animation movie, and 2.0 for the natural movie sequence. This shows that the users are satisfied with the quality of their preferred video streams. In summary it can be observed that potential users of mobile devices prefer a higher number of details visible per frame and therefore also accept a lower frame rate of the stream. Only in the case that there are very many scene cuts within the sequence, a higher frame rate was preferred over a more detailed version to get a smoother motion between those cuts.

C. Detail Adaptation

The results from the user interviews concerning their preferred frame rate clearly indicate that a high detail resolution is a crucial factor for the users of mobile devices. The detail resolution of a video directly relates to the bit rate of the video stream which in turn is limited by the rate that the client is able to receive from the network. Thus, the detail resolution needs to be reduced as much as required from the network connection of the client.

Similar as in the discussion about the other dimensions the question arises if the detail resolution can be reduced even further while the quality is acceptable by the users of mobile devices. In order to inspect if there is a certain quality level at which users cannot notice any further quality enhancements on a mobile device, we conducted further subjective quality tests. We encoded one of the previously used sequences at different bit rates and presented them pairwise to potential users. Afterwards, the test persons were asked to decide which version of the presented videos they liked more. Because the test persons may have chosen the preferred sequence randomly, they were also asked if they really noticed the better quality or just guessed.

		Version A		
		500 kbit/s	700 kbit/s	1500 kbit/s
	300 kbit/s	87.80 %	85.37 %	90.24 %
Version B	500 kbit/s		70.73 %	73.17 %
	700 kbit/s			60.98 %

TABLE I
PERCENTAGES OF PEOPLE PREFERRING VERSION A OVER B

Four versions with different bit rates result in six pairs of videos with different bit rates. 41 non-expert users took part in these tests. Table I shows the relative frequencies of the sequences which quality was rated better in each pairwise comparison between a version A and a version B of the same video sequence. Each column in this table contains the portion of users which preferred the version with the bit rate given in the first row to the version with the bit rate given in the corresponding row. The version with a bit rate of 700 kbit/s, for instance, was chosen to have a better quality than the version with 500 kbit/s by 70.73 % of the users. Compared to the version with 1500 kbit/s, however, only 39.02 % of the users said that the quality of the 700 kbit/s version was better.

These results show that there is a clear preference for those versions of the soccer sequences with a higher bit rate and therefore also a higher detail resolution. For all six comparisons together, 78.04 % of the users chose the sequence with the higher bit rate.

In 23.57 % of all comparisons the users stated that they did not notice any differences between the two versions of the video and that they chose one version randomly. In 60.34 % of these cases, however, the users intuitively chose the version with a higher bit rate to have higher quality. This high percentage shows that we can assume that these choices were not truly by pure chance. Although the participants stated that they are not aware of any differences, the results show that the higher quality is still noticeable by the users. Thus, for this video sequence, there is no optimal quality level at a reasonable rate up to 1500 kbit/s.

D. Combined Adaptation

In our previous work we inspected the impact of different video adaptation dimensions on the requirements of the client. We have further identified three major dimensions which need to be adapted in order to support a great heterogeneity of mobile devices: the spatial, the temporal, and the detail resolution. For these dimensions we analyzed how video adaptation in these dimensions may affect the quality of the produced video streams. In order to support a great heterogeneity of devices, however, a combined adaptation of the spatial, the temporal as well as the detail resolution is needed. To get an idea of how an adaptation of those different dimensions affects the bit rate of the stream, we re-encoded the test sequences that we had used before in four different versions:

i) adapted in the detail dimension
ii) adapted in the temporal and the detail dimension
iii) adapted in the spatial and the detail dimension
iv) adapted in the spatial, temporal and detail dimension

Figure 7 exemplarily shows the bit rates of the mobile sequence for all four different versions. Apart from the absolute values, there are only small differences between the graphs for all inspected sequences, although they differ greatly in terms of the amount of motion.

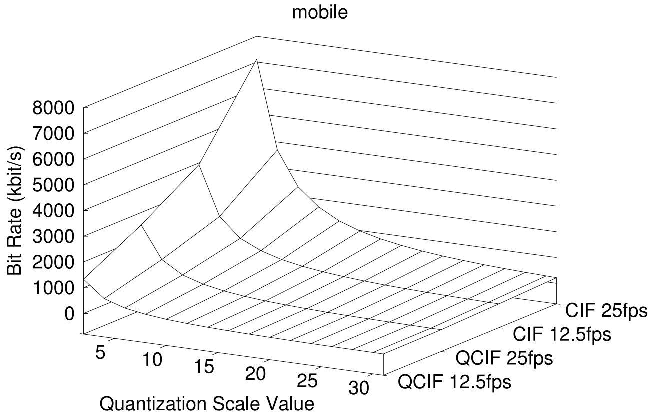

Fig. 7. Video bit rate at different spatial, temporal and detail resolutions

The existing differences in the graphs of all test sequences can be traced back to the amount of motion in the stream. The akiyo sequence, for instance, has a very low amount of motion and thus, the inter-coded frames consume only a low amount of the bit rate compared to the intra-coded frames because the motion prediction produces only very small residual errors that need to be encoded. When the frame rate is reduced, the inter-coded frames need to carry much information as the differences between the frames are increasing. In case of the akiyo sequence the size of the P-frames is increased by 41.31 % when the frame rate is reduced from 25 to 12.5 fps. Therefore, the bit rate of the stream is reduced just to 70.13 % although the frame rate is reduced to 50 %. The mobile sequence contains a high amount of motion and therefore, the inter-coded frames already carry a high residual error. The size of the P-frames is increased by the frame rate reduction only by 10.64 % and therefore, the bit rate of the stream is reduced to 55.35 % due to the reduction of the frame rate.

A reduction of the spatial resolution by a factor of two in both dimensions reduces the bit rate of the stream to about 36 % for the mobile sequence in the case of the lowest quantizer scale value. For other sequences, similar reductions can be found and thus, the different amount of motion does not significantly influence the amount of reduction. The reason that the bit rate is not reduced to a fourth of the original bit rate results from the motion information contained in the frames. The graph also shows that the amount of bit rate reduction decreases for higher quantizer scale values.

III. Conclusion

In this paper we present the results of our investigations concerning the impact of three different adaptation dimensions, i.e., the temporal, the spatial as well as the detail dimension, on the produced quality of video stream. For the spatial dimension we encoded several test sequences at different spatial resolutions and compared their PSNR values. Although the values of the streams at the target resolution were slightly lower than the values at the original resolution of the stream, an adaptation to the target resolution should be preferred, due to the processing overhead needed for downscaling the stream to the target resolution at the client otherwise. For the temporal as well as for the detail resolution we conduced some user interviews. The users accepted a lower frame rate in order to get more visible details. However, we could not identify an upper bound for the detail resolution at a reasonable bit rate. Additionally, we also investigated the produced bit rate when adapting the streams in all three dimensions.

In summary, we can conclude that an optimal adaptation should firstly tailor the spatial resolution of the stream to the display resolution of the requesting client. This also reduces the bit rate of the stream significantly and further reduction can be achieved by reducing the temporal resolution. Finally, the detail resolution can be reduced as needed to fine tune the bit rate of the stream to the network connection of the client.

References

[1] J. Brandt and L. Wolf, "Multidimensional Transcoding for Adaptive Video Streaming," in *Proceedings of the 17th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'07)*, Urbana-Champaign, IL, USA, Jun. 2007, pp. 57–62.

[2] J. D. McCarthy, M. A. Sasse, and D. Miras, "Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video," in *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '04)*, Vienna, Austria, Apr. 2004, pp. 535–542.

[3] A. Eichhorn and P. Ni, "Pick your Layers wisely - A Quality Assessment of H.264 Scalable Video Coding for Mobile Devices," in *IEEE International Conference on Communications, ICC '09*, Dresden, Germany, Jun. 2009.

[4] M. Li, M. Claypool, R. Kinicki, and J. Nichols, "Characteristics of streaming media stored on the Web," *ACM Transactions on Internet Technology (TOIT)*, vol. 5, no. 4, pp. 601–626, Nov. 2005.