Content-Based Adaptation of Streamed Multimedia

Nikki Cranley[1], Liam Murphy[1], and Philip Perry[2]

[1] Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland
{Nicola.Cranley,Liam.Murphy}@UCD.ie
[2] School of Electronic Engineering, Dublin City University,
Glasnevin, Dublin 9, Ireland
PerryP@eeng.DCU.ie

Abstract. Most adaptive delivery mechanisms for streaming multimedia content do not explicitly consider user-perceived quality when making adaptations. We show that an Optimal Adaptation Trajectory (OAT) through the set of possible encodings exists, and that it indicates how to adapt encoding quality in response to changes in network conditions in order to maximize user-perceived quality. The OAT is related to the characteristics of the content, in terms of spatial and temporal complexity. We describe an objective method to automatically determine the OAT in response to the time-varying characteristics of the content. The OAT can be used with any transmission adaptation policy. We demonstrate content-based adaptation using the OAT in a practical system, and show how this form of adaptation can result in differing adaptation behaviour.

1 Introduction

Best-effort IP networks, particularly wireless networks, are unreliable and unpredictable. There can be many factors that affect the quality of a transmission, such as delay, jitter and loss. Adaptation techniques should attempt to reduce network congestion and packet loss by matching the rate of the video stream to the available network bandwidth. Without adaptation, any data transmitted exceeding the available bandwidth could be discarded, lost or corrupted in the network. This has a devastating effect on the playout quality of the received stream. A slightly degraded quality but uncorrupted video stream is less irritating to the user than a corrupted stream. In general, adaptation policies (whether sender-based [1], receiver-based [2],[3], or encoder-based are [4]) address the problem of how to adapt only in terms of adjusting the transmission rate or the window size and are thus bitrate centric. Other adaptation approaches include utility-based schemes [5],[6], which adapt video quality encoding configurations by using a utility function (UF). However, rapidly fluctuating quality should also be avoided as the human vision system (HVS) adapts to a specific quality after a few seconds and it becomes annoying if the viewer has to adjust to a varying quality over short time scales [7]. Controlled video quality adaptation is needed to reduce the negative effects of congestion on the stream whilst providing the highest possible level of service and perceived quality.

In previous work we proposed that there is an optimal way in which multimedia transmissions should be adapted in response to network conditions to maximize the

J. Vicente and D. Hutchison (Eds.): MMNS 2004, LNCS 3271, pp. 39–49, 2004.

user-perceived quality. Extensive subjective testing demonstrated the existence of an Optimum Adaptation Trajectory (OAT) in the space of possible encodings and that it is related to the content type [8]. However, due to the time-varying nature of content characteristics, there is a need to automatically and dynamically determine the OAT based on these contents characteristics in order to properly apply the OAT. This knowledge can then be used as part of a content-based adaptation strategy, which aims to maximize the user-perceived quality of the delivered multimedia content.

This paper is structured as follows. Section 2 describes the concept of an Optimum Adaptation Trajectory (OAT) that exists for each class of video content. Section 3 describes how content can be characterized by its spatial and temporal complexities. Section 4 presents an objective means of determining the OAT that is dynamic and can react to the time varying characteristics of content. Section 5 describes our content-based adaptive streaming system. The system is demonstrated using the existing Loss-Delay Adaptation (LDA) algorithm using a content-based dynamic OAT. Some preliminary simulation results from our system are presented to show system operation and behavior. Conclusions and directions for future work are presented in Section 6.

2 Optimum Adaptation Trajectories

In previous work, the concept of an Optimum Adaptation Trajectory (OAT) has been presented [8] and has been shown to complement the sender-based Loss-Delay Adaptation algorithm using a static OAT determined by subjective testing, or as the basis of a Perceptual Quality Adaptation (PQA) algorithm [9].

The OAT embodies the idea that there is an optimal way in which multimedia transmissions should be adapted (upgraded/downgraded) in response to network conditions to maximize the user-perceived quality. This is based on the hypothesis that within the set of different ways to achieve a target bit rate, there exists an encoding configuration that maximizes the user-perceived quality. If a particular multimedia file has n independent encoding configurations then, there exists an adaptation space with n dimensions. Adaptation space consists of all possible dimensions of adaptation for the content that can be implemented as part of an adaptive streaming server or adaptive encoder. When adapting the transmission from some point within that space to meet a new target bit rate, the adaptive server should select the encoding configuration that maximizes the user-perceived quality for that given bit rate. The example shown in Figure 1 indicates that, when degrading the quality from an encoding configuration of 25fps and a spatial resolution of 100%, there are a number of possibilities – such as reducing the frame rate only to 15fps, reducing the spatial resolution only to 70%, or reducing a combination of both the frame rate and resolution. The choice of which encoding configuration that should be adopted is determined as the encoding configuration that maximizes the user-perceived quality. When the transmission is adjusted across its full range, the locus of these selected encoding configurations should yield an OAT within that adaptation space.

There is much research into developing objective metrics for video quality assessment [10],[11],[12],[13]. The most commonly used objective metric of video quality

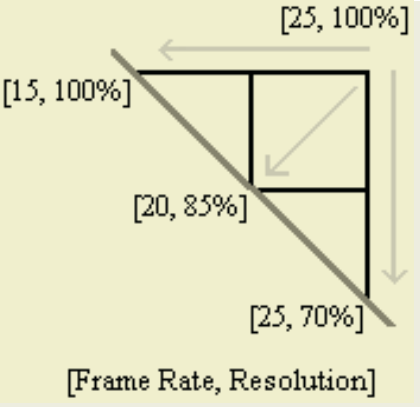

Fig. 1. Encoding Quality Options.

assessment is the Peak Signal to Noise Ratio (PSNR), which has been widely used in many applications and adaptation algorithms [14] to assess video quality. The advantage of PSNR is that it is very easy to compute using the Mean Square Error (MSE) of pixel values of luminance for frames from the degraded and reference clips. However, PSNR does not match well to the characteristics of the human vision system (HVS) [15]. However, the main problem with using PSNR values as a quality assessment method is that even though two images may be different, the visibility of this difference is not considered. The PSNR metric does not take into consideration any details of the HVS such as its ability to "mask" errors that are not significant to the human comprehension of the image. Several objective metrics, namely the Video Quality Metrics (VQM) and PSNR were investigated to determine whether they yielded an OAT. The ITU-T has recently accepted the VQM as a recommended objective video quality metric that correlates adequately to human perception in [16], [17], [18]. However, our findings were that these objective metrics produce an OAT which jumped through adaptation space with no sense of direction or continuity. In contrast, subjective methods were consistent across content types and produced a smooth graceful OAT through adaptation space.

The OAT is dependent on the characteristics of the content. There is a content space in which all types of video content exist in terms of spatial and temporal complexity (or detail and action). The OAT was discovered through extensive subjective testing for a number of different content types using the forced choice methodology. Subjective testing results showed that the OAT is logarithmic with the general form

$$Resolution = A*Ln(Frame\ rate) + B \quad (1)$$

for some constants A and B. It was also found that the temporal and spatial complexity of the scene plays an important role in the curvature of the OAT. The usefulness of the OAT relies on the contents' spatial and temporal characteristics being known by the adaptive server. Since the spatial and temporal complexity of content will vary over time, we propose a method to automate the process of determining the OAT in response to these changing content characteristics.

3 Spatial and Temporal Complexity Metrics

User perception of video quality varies with the content type; for example, viewers perceive action clips differently from slow moving clips. Thus, there exists a different OAT for different types of content based on their spatial and temporal characteristics. The spatial and temporal complexity of content can be determined using the metrics Spatial Information (SI) and Temporal Information (TI).

The Spatial Information parameter, SI, is based on the Sobel filter, and is implemented by convolving two 3 x 3 kernels over the luminance plane in the video frame [19]. Let $Conv1_n(i, j)$ be the result of the first convolution for the pixel of the input nth frame at the ith row and jth column and let $Conv2_n(i, j)$ be the result of the second convolution for the same pixel. The output of the Sobel filtered pixel at the ith row and jth column in the nth frame, $y_n(i,j)$, is the square root of the sum of the squares of both convolutions. The SI value is the standard deviation (std_{space}) over all pixels in the nth frame and is computed as follows:

$$y_n(i, j) = \sqrt{[Conv1_n(i, j)]^2 + [Conv2_n(i, j)]^2} \tag{2}$$

$$SI = std_{space}[y_n] \tag{3}$$

This process is repeated for each frame in the video sequence and results in a time series of spatial information of the scene. The calculations are performed on a sub-image of the video frame to avoid unwanted edge and border effects. The size of the original image is QCIF (176x144 pixels) and so a centrally located sub-image of 100x100 was used.

The Temporal Information parameter, TI, is based upon the motion difference feature in successive frames. The motion difference, $M_n(i, j)$, is the difference between the pixel values in the ith row and jth column in nth frame $F_n(i, j)$, and the value for the same pixel in the previous frame, $F_{n-1}(i, j)$. The measure of Temporal Information, TI, is the standard deviation over all pixels in the sub-image space (std_{space}) and is computed as follows:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j) \tag{4}$$

$$TI = std_{space}[M_n] \tag{5}$$

Both the SI and TI values result in a time varying measure of the spatial and temporal complexity of a piece of content. This information can be used to create a time varying understanding of the characteristics and requirements of the content to generate the OAT. The calculation of the SI and TI complexity parameters is not computationally expensive for small image sizes such as QCIF.

4 Objective OAT

Upon analysis of the OATs discovered by subjective testing, it was observed that the OAT for different content types was not strongly dependent on the precise SI-TI values, but more influenced by the relationship between spatial and temporal com-

plexity. For example, when the SI value was significantly greater than the TI value, the resulting OAT tended towards the spatial resolution, and vice versa. However, for a number of test sequences where the spatial and temporal complexities were approximately equal, the OATs were "neutral".

To represent the relative dominance of one characteristic over another, a weighting factor, W, is introduced which is determined using the SI and TI metrics. The factor W is the relative dominance of temporal complexity over the spatial complexity. Since the scales of spatial complexity and temporal complexity are different, both parameters were converted to their respective fractional values. The fractional SI value is thus the SI value divided by the maximum SI value; similarly, the fractional TI value is the TI value divided by the maximum TI value. The maximum SI and TI values can be found by applying the equations to the luminance plane of an image with alternating black and white pixels.

$$Fractional\_SI = SI / SI_{MAX} \tag{6}$$

$$Fractional\_TI = TI / TI_{MAX} \tag{7}$$

$$\mathrm{W} = \text{Weighting factor} = \left(Fractional\_TI / Fractional\_SI \right) \tag{8}$$

From Figure 2, it can be seen that when SI=TI, the weighting factor, W, is equal to 1, therefore there is no dominant characteristic and the OAT is neutral. If SI>TI, then the weighting factor W<1 and the spatial complexity is dominant, and the resulting OAT should tends towards maintaining the spatial resolution during adaptation. Conversely, when SI<TI, the weighting factor W>1 and the resulting OAT should tend towards maintaining the frame rate during adaptation. The following empirical equation (Eqn. 9) was derived to relate the OATs discovered by subjective testing, the weighting factor, W, resolution and frame rate:

$$\mathrm{Re}\, s = \left(\mathrm{Re}\, s_{MAX} / Ln(F_{MAX}) \right) \left(WLn(F) - (W-1)Ln(F_{MAX}) \right) \tag{9}$$

where:

Res = Spatial resolution;
Res_{MAX} = Maximum spatial resolution = 100%;
F_{MAX} = Maximum frame rate = 25fps;
F = Frame rate.

From Figure 3 it can be seen that the OAT increases in curvature towards the frame rate with increasing values of W. For very low values of W only the spatial resolution should be degraded: this would be expected for film credits or panoramic still shots, where there is very low temporal information but high spatial resolution requirements. The objective OATs demonstrate that a piece of content containing a static still image (TI value of zero) should be adapted in the frame rate dimension only. However, it is not possible for a piece of content to have a zero SI value and a non-zero TI value.

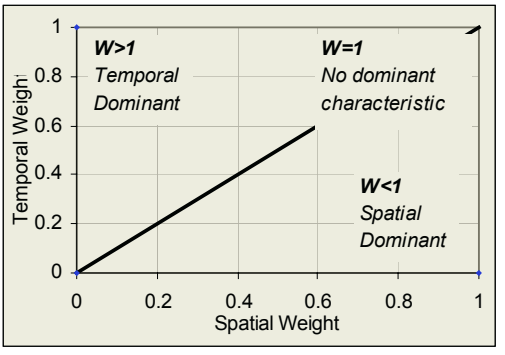

Fig. 2. Spatial-Temporal Space.

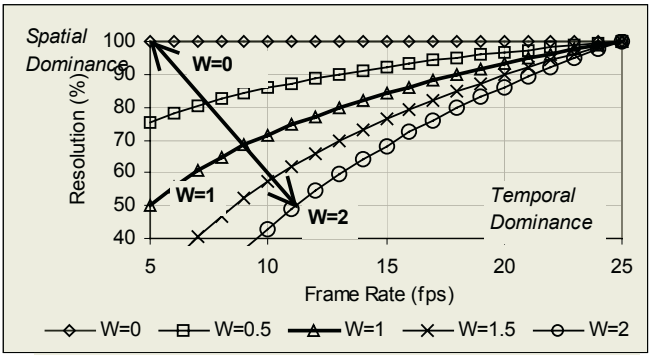

Fig. 3. Objective OAT variance with different weighting factors.

The objective OAT was validated by comparing the OAT discovered by subjective testing and that determined using the objective OAT. The SI and TI values for several different content types were measured to determine the weighting factor. From this, the objective OAT was calculated and then compared against the OAT discovered by subjective testing. The results in Figure 4 indicate a high degree of correlation between the objective OAT and the subjective OAT, which, in most cases was over 98%.

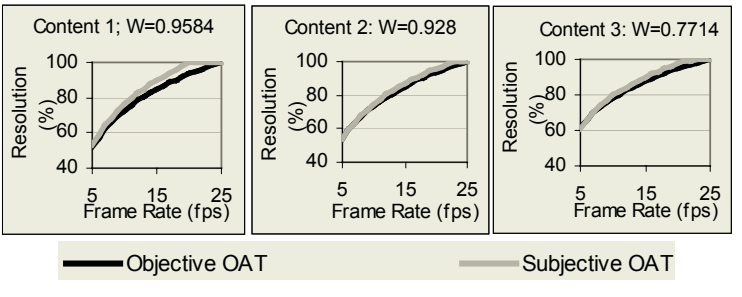

Fig. 4. Correlation of Subjective and Objective OATs.

5 Content-Based Adaptation

Dynamic content-based adaptation can be integrated into a client server system with either in a real-time system with real-time encoding and analysis (Figure 5) or else for streaming of pre-analyzed pre-encoded content (Figure 6). Our prototype system is of the latter architecture. Both client and server consist of the RTP/UDP/IP stack with RTCP/UDP/IP to relay feedback messages between the client and server. The content is pre-encoded and hinted with an MPEG-4 encoder. The content contains multiple tracks, each encoded with a different resolution. By switching tracks the server can dynamically and discretely adapt the resolution. The server can also apply a frame dropping policy to dynamically adapt the frame rate. By controlling these two operations the server is able to implement the two-dimensional adaptation trajectory given by the OAT.

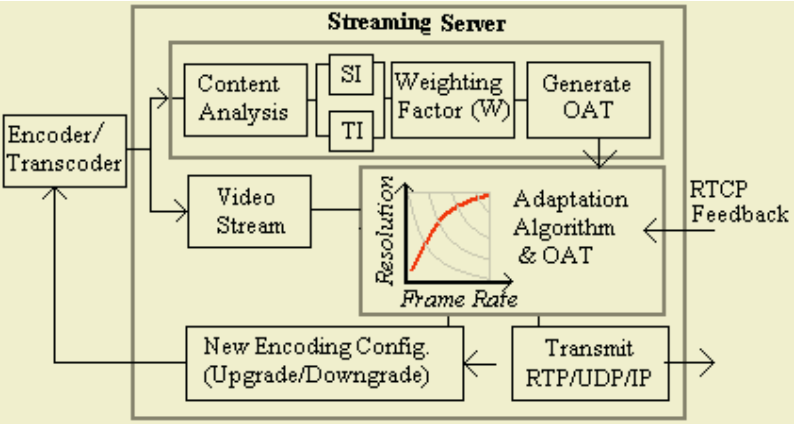

Fig. 5. Basic System Architecture for Live Content.

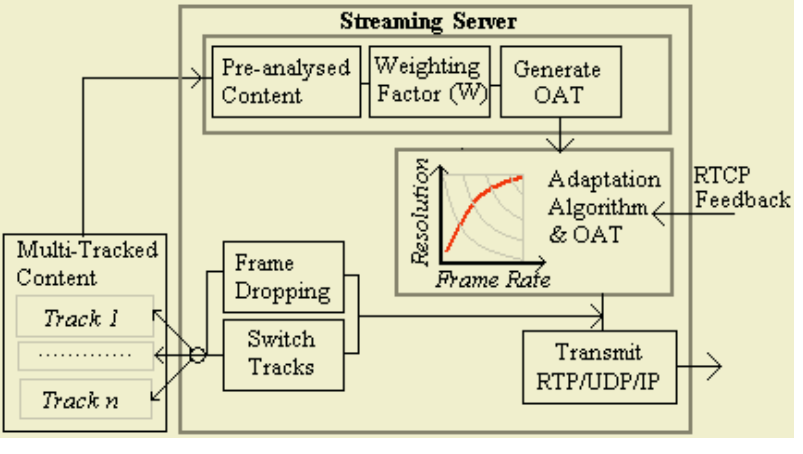

Fig. 6. Basic System Architecture for Pre-encoded Content.

To demonstrate how the objective OAT can be used to provide content-based adaptation to complement the LDA adaptation algorithm, a prototype client-server system was developed. The client returns standard RTCP-RR feedback containing information about loss, delay and bottleneck bandwidth values. When the server receives feedback from the client, the LDA algorithm indicates how the bit rate should be adjusted in response to fluctuations in the available end-to-end bit rate between client and server. The server finds the intersection of the OAT and the new target bit rate as determined by the adaptation algorithm. From this intersection point of the new target bit rate and the OAT, the server finds the corresponding encoding configuration on the OAT indicating the quality-encoding configuration that maximizes the user-perceived quality for the content. Having found this new encoding configuration, the server adjusts the frame rate and/or adapts the resolution by switching tracks. The OAT is constantly varied in response to the changing characteristics of the content.

Given that the weighting factor of the content typically changes in a subtle and gradual manner (with the exception of scene changes), the weighting factor was averaged over 20 second intervals. The time variance of the weighting factor can be seen in Figure 7(a). A challenging network condition has been selected to demonstrate the efficacy and use of content-based adaptation using the OATs with LDA. This would be typical of a wireless IP network where mobility can result in sudden and substantial changes in the available bit rate. In the simulations, RTCP feedback was fixed at every 5 seconds. The stability of our system and its ability to react to network conditions is entirely dependent on the frequency of feedback as the system can adapt on each feedback report. In the example below, Figure 7(b) shows how the server's

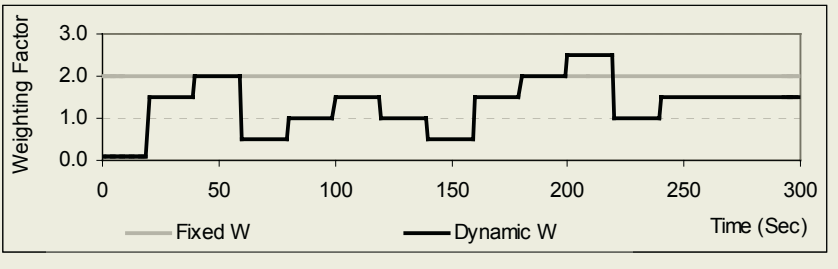

Fig. 7a. Weight factor variations with time.

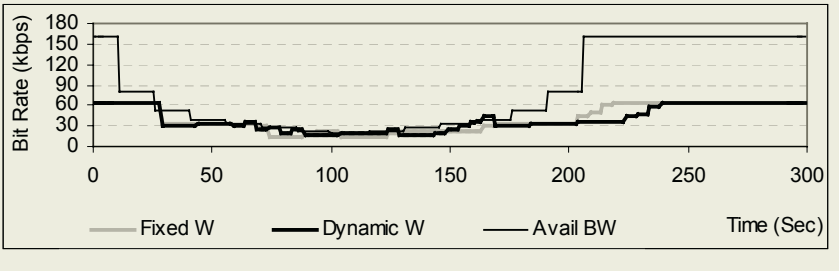

Fig. 7b. Bit rate variations with time.

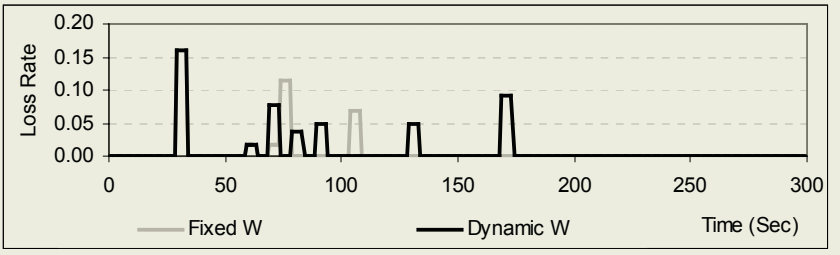

Fig. 7c. Loss rate variations with time.

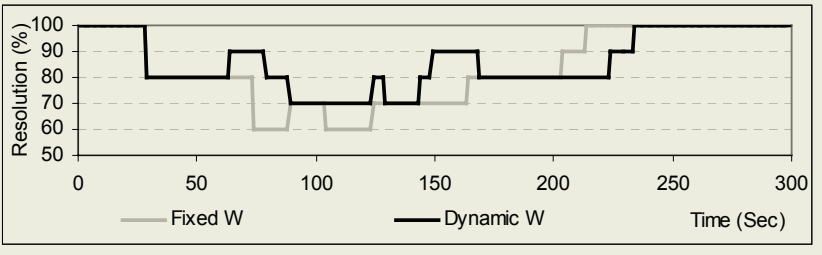

Fig. 7d. Resolution variations with time.

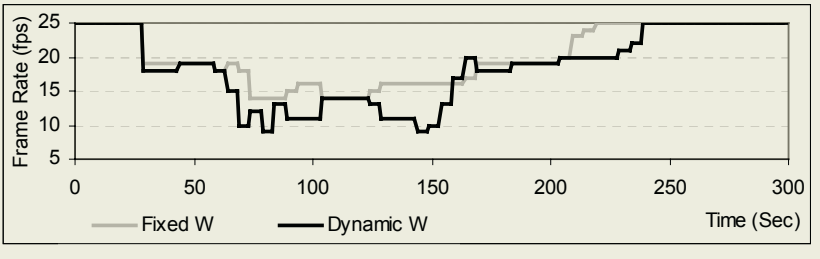

Fig. 7e. Frame rate variations with time.

transmission rate (using both a fixed weighting factor and a dynamic weighting factor) adapts in response to the available bandwidth ("Avail BW") at the client using the LDA algorithm whilst Figure 7(c) shows the loss rate. The intersection of the target transmission rate given by LDA and the intersection with the OAT gives the required resolution and frame rate. Figures 7(d) and Figure 7(e) show how the resolution and frame rates are adapted. Adaptation occurs in both dimensions of frame rate and resolution simultaneously as indicated by the OAT. These examples show the very different behaviors that can result from content-based adaptation. When the weighting factor is low, the resolution is increased faster during periods of no congestion and decreased at a slower rate when congestion occurs. By using an automatically generated OAT that is related to the characteristics of the content, it is expected that this would enhance the user perception and quality of the session since the quality is degraded and upgraded in a known and controlled manner that has the least

negative impact on the perceptual quality of the content and is based on the characteristics of the content.

6 Conclusions and Future Work

We have built upon previous work in which the concept of an Optimum Adaptation Trajectory was proposed and shown to exist by subjective testing. The OAT proposed that there is an optimal way in which multimedia transmissions should be adapted (upgraded/downgraded) in response to network conditions to maximize the user-perceived quality. The OAT is related to the spatial and temporal characteristics of the content or more specifically the relative dominance of one characteristic over another. In this paper we have indicated how the characteristics of the content can be encapsulated by a weighting factor. This weighting factor plays an important role in the empirically derived equation used to generate the OAT. This equation was shown to correlate well to the OATs discovered by subjective testing. Finally, we showed how the dynamic content-based OAT can be used with the sender-based LDA algorithm.

Future work involves integrating content-based adaptation into an adaptation algorithm that uses the OAT directly as a means of adaptation. Further subjective testing is required to verify overall better user-perceived quality using content-based adaptation. It is possible to increase the feedback frequency and work is underway to implement and investigate the effects and efficacy of increased feedback frequency, as proposed by the 3GPP organization who suggest using a bandwidth modifier in the Session Description Protocol (SDP) to increase the RTCP feedback frequency [20] such that an RTCP feedback packet is sent at least once a second [21].

Acknowledgements

The support of the Research Innovation Fund and Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

References

1. D. Sisalem, A. Wolisz LDA+: A TCP-friendly adaptation scheme for multimedia communication, IEEE International Conference on Multimedia and Expo (III) (2000).
2. D. Sisalem, A. Wolisz, MLDA: A TCP-friendly congestion control framework for heterogeneous multicast environments, Proc. Eighth International Workshop on Quality of Service (IWQoS 2000), Pittsburgh, PA, June (2000).
3. V. Jacobson S. McCanne, M. Vetterli. Receiver-driven layered multicast, Proc. of ACM SIGCOMM'96, Stanford, CA, August (1996).
4. D. Wu, Y. T. Hou, W. Zhu, H.-J. Lee, T. Chiang, Y.-Q. Zhang, H. J. Chao, On end-to-end architecture for transporting MPEG-4 video over the Internet, IEEE Trans. on Circuits and Systems for Video Technology, vol. 10, no. 6, Sept. (2000)
5. J.G. Kim, Y. Wang, S.F. Chang, Content-adaptive utility-based video adaptation, IEEE ICME 2003, Baltimore, July (2003)

6. Y. Wang, J.G. Kim, S.F. Chang, Content-based utility function prediction for real-time MPEG-4 transcoding, ICIP 2003, Barcelona, Spain, September 14-17, (2003).
7. G. Ghinea, J.P. Thomas, R. Fish, Multimedia, Network Protocols and Users - Bridging the Gap, ACM Multimedia '99, pp. 473-476, Orlando, Florida, (1999)
8. N.Cranley, L. Murphy, P. Perry, User-Perceived Quality Aware Adaptive Delivery of MPEG-4 Content, Proc. NOSSDAV'03, Monterey, California, June (2003)
9. N.Cranley, L. Murphy, P. Perry, Perceptual Quality Adaptation (PQA) algorithm for 3GP and multi-tracked MPEG-4 content over wireless networks, Proc. 15th IEEE Intl. Symp. on Personal, Indoor and Mobile Radio Communications, Barcelona, Spain, Sept. (2004)
10. M.J. Nadenau, S. Winkler, et al., Human Vision Models for perceptually Optimized Image Processing - A Review, Proceedings of the IEEE 2000, (2000)
11. S. Winkler, Vision Models and Quality Metrics for Image Processing Applications, Ph.D. Thesis, Ecole Polytechnique Federale de Lausanne (EPFL), (2000)
12. Z. Yu, H.R. Wu, Human visual system based objective digital video quality metrics, in Proc. Intl. Conf. on Signal Processing of IFIP World Computer Conference 2, August (2000)
13. M. Masry, S.S Hemami, Models for the perceived quality of low bit rate video, Proc. of IEEE International Conference on Image Processing, Rochester, NY, September (2002)
14. J.G. Kim, Y. Wang, S.F. Chang, Content-adaptive utility-based video adaptation, IEEE ICME 2003, Baltimore, July (2003)
15. C. Van den Branden Lambrecht, O. Verscheure, Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System, Proceedings of SPIE 96, San Jose, CA, (1996)
16. Institute for Telecommunication Sciences, Technical Report, 2002, ITU-T and Related U.S. Standards Development, http://its.bldrdoc.gov/tpr/2002/itu_related_standards.pdf, (2002)
17. ITU-T Recommendation J.149, Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM), (2004)
18. ITU-T Recommendation J.148, Requirements for an objective perceptual multimedia quality model, (2003)
19. ITU-T P.910 Recommendation, Subjective video quality assessment methods for multimedia applications, (1996)
20. 3GPP TSG-SA WG4, Tdoc S4-030019, RTCP Packet Frequency for very low bit rate sessions, (2004)
21. S. Casner, SDP Bandwidth Modifier for RTCP Bandwidth, <draft-ietf-avt-rtcp-bw-05.txt>, (2002)