Content Clustering Based Video Quality Prediction Model for MPEG4 Video Streaming over Wireless Networks

Asiya Khan, Lingfen Sun and Emmanuel Ifeachor
Centre for Signal Processing and Multimedia Communication
School of Computing, Communications and Electronics
University of Plymouth, Plymouth PL4 8AA, UK.
Email: asiya.khan@plymouth.ac.uk; L.Sun@plymouth.ac.uk; E.Ifeachor@plymouth.ac.uk

Abstract— **The aim of this paper is quality prediction for streaming MPEG4 video sequences over wireless networks for all video content types. Video content has an impact on video quality under same network conditions. This feature has not been widely explored when developing reference-free video quality prediction model for streaming video over wireless or mobile communications. In this paper, we present a two step approach to video quality prediction. First, video sequences are classified into groups representing different content types using cluster analysis. The classification of contents is based on the temporal (movement) and spatial (edges, brightness) feature extraction. Second, based on the content type, video quality (in terms of Mean Opinion Score) is predicted from network level parameter (packet error rate) and application level (i.e. send bitrate, frame rate) parameters using Principal Component Analysis (PCA). The performance of the developed model is evaluated with unseen datasets and good prediction accuracy is obtained for all content types. The work can help in the development of reference-free video prediction model and priority control for content delivery networks.**

***Index Terms*—MOS, Content clustering, Video quality prediction, Video streaming**

I. INTRODUCTION

Streaming video services are becoming commonplace with the recent success of broadband access lines. Perceived quality of the streaming video is likely to be the major determining factor in the success of the new multimedia applications. It is therefore important to choose or adapt both the application level i.e. the compression parameters as well as network settings so that maximum end-to-end user perceived video quality can be achieved. Reference-free video quality prediction model can be used for on-line video quality monitoring and assisting video quality adaptation/control for maximizing video quality.

The prime criterion for the quality of multimedia applications is the user's perception of service quality [1]. The most widely used metric is the Mean Opinion Score (MOS). Video quality is dependent on the type of content e.g. network requirement for contents like sports is higher compared to that of news. Furthermore, video transmission over wireless networks is highly sensitive to transmission problems such as packet loss or network delay. Several objective metrics for perceptual video quality estimation have been proposed recently. In [2],[3] authors propose an opinion and parametric model for estimating the quality of interactive multimodal and videophone services that can be used for application and/or network planning and monitoring. However, in these work content types are not considered. In [4] a theoretical framework is proposed based on both application and network level parameters to predict video quality. Work in [5] is only based on network parameters. (e.g. network bandwidth, delay, jitter and loss) to predict video quality with no consideration of application-level parameters. In [6] we have proposed an ANFIS-based prediction model that considers both application and network level parameters with subjective content classification. Recent work has also shown the importance of video content in predicting video quality. In [7],[8] video content is classified based on the spatial (edges, colours, etc) and temporal (movement, direction, etc) feature extraction which were then used to predict video quality together with other application-level parameters such as send bitrate and frame rate. However, this work did not consider any network-level parameters in video quality prediction. Video content is classified in [9],[10] based on content characteristics obtained from users' subjective evaluation using cluster [11] and Principal Component Analysis (PCA) [12]. In [13],[14] authors have used a combination of PCA [12] and feature extraction to classify video contents.

In this paper, first we aim to recognize the most significant content types, classify them into groups using a combination of temporal (movement) and spatial (edges, brightness) feature extraction using a well known tool called cluster analysis [11]. Secondly, to develop a reference-free video prediction model at the user level (perceptual quality of service) in terms of the MOS for all content types combining both the application level (send bitrate, frame rate) and network level (packet error rate) parameters. Our focus ranges from low resolution and low send bitrate video streaming for 3G applications to higher video send bitrate for WLAN applications depending on type of content and network conditions. The proposed test bed is based on simulated network scenarios using a network simulator (NS2) [15] with an integrated tool Evalvid [16]. It gives a lot of flexibility for evaluating different topologies and parameter settings used in this study.

The paper is organized as follows. Section II outlines the evaluation set-up. In section III we classify the contents based on cluster analysis. Section IV predicts the video quality,

whereas, section V concludes the paper and highlights areas of future work.

II. Evaluation Set-Up

For the tests we selected nine different video sequences of qcif resolution (176x144) and varied duration from 10 to 30s. The video sequences were encoded in MPEG4 format with an open source ffmpeg [17] encoder/decoder with a Group of Pictures (GOP) pattern of IBBPBBPBB. Each GOP encodes three types of frames - Intra (I) frames are encoded independently of any other type of frames, Predicted (P) frames are encoded using predictions from preceding I or P frames and Bi-directionally (B) frames are encoded using predictions from the preceding and succeeding I or P frames

A GOP pattern is characterized by two parameters, GOP(N,M) – where N is the I-to-I frame distance and M is the I-to-P frame distance. For example, as shown in Fig.1, G(9,3) means that the GOP includes one I frame two P frames, and six B frames. The second I frame marks the beginning of the next GOP. Also the arrows in Fig. 1 indicate that the B frames and P frames decoded are dependent on the preceding or succeeding I or P frames [18].

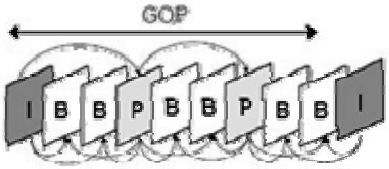

Fig. 1 A sample of MPEG4 GOP (N=9, M=3)

The chosen video sequences ranged from very little movement, i.e. small moving region of interest on static background to fast moving sports clips. Each of the test sequences represent typical content offered by network providers.

TABLE I
TESTBED COMBINATIONS

Video sequences	Frame Rate fps	SBR (kb/s)	PER
Akiyo	10, 15, 30	18	0.01, 0.05, 0.1, 0.15, 0.2
Suzie	10, 15, 30	44	
Grandma	10, 15, 30	80	
Carphone	10, 15, 30	44	0.01, 0.05, 0.1, 0.15, 0.2
Foreman	10, 15, 30	80	
Rugby	10, 15, 30	128	
Stefan	10, 15, 30	104	0.01, 0.05, 0.1, 0.15, 0.2
Table tennis	10, 15, 30	384	
Football	10, 15, 30	512	

For quality evaluation we used a combination of application and network level parameters as Frame Rate (FR), Send Bitrate (SBR) and Packet Error Rate (PER). The video sequences along with the combination parameters chosen are given in Table I. In total, there were 1500 encoded test sequences.

To obtain a MOS (Mean Opinion Score) value we conducted experiments with an open source framework Evalvid [16] and network simulator tool NS2 [15]. Video quality is measured by taking the average PSNR (Peak-Signal-to-Noise-Ratio) over all the decoded frames. MOS scores are calculated objectively based on PSNR to MOS conversion from Evalvid [16] given in Table II below. Future work will involve taking subjective MOS scores to validate our results.

TABLE II
PSNR TO MOS CONVERSION

PSNR (dB)	MOS
>37	5
31 – 36.9	4
25 – 30.9	3
20 – 24.9	2
< 19.9	1

The experimental set up is given in Fig 2. There are two sender nodes as CBR background traffic and MPEG4 video source. CBR background traffic is added to make the simulation results more realistic. Both the links pass traffic at 10Mbps, 1ms over the internet which in turn passes the traffic to another router over a variable link. The second router is connected to a wireless access point at 10Mbps, 1ms and further transmits this traffic to a mobile node at a transmission rate of 11Mbps 802.11b WLAN. No packet loss occurs in the wired segment of the video delivered path. The maximum transmission packet size is 1024 bytes. The video packets are delivered with the random uniform error model. The CBR rate is fixed to 1Mbps to give a more realistic scenario. The packet error rate is set in the range of 0.01 to 0.2 with 0.05 intervals. To account for different packet loss patterns, 10 different initial seeds for random number generation were chosen for each packet error rate. All results generated in the paper were obtained by averaging over these 10 runs.

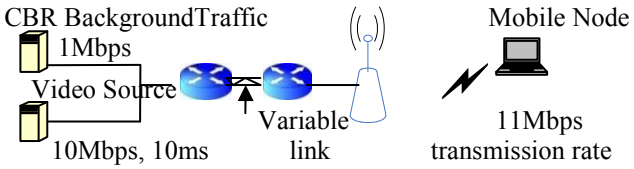

Fig. 2 Simulation setup

III. Content Classification Based On Cluster Analysis

Video content is classified using a well known multivariate statistical analysis called cluster analysis [11]. This technique is used as it groups samples that have various characteristics into similar groups. Cluster analysis is carried out on the nine video sequences given in Table I based on the temporal and spatial feature extraction.

A. Temporal feature extraction

The movement in a video clip given by the SAD value (Sum of Absolute Difference). The SAD values are computed as the pixel wise sum of the absolute differences between the two frames being compared and is given by:

$$\mathrm{SAD}_{n,m} = \sum_{i=1}^{N}\sum_{j=1}^{M}|B_n(i,j) - B_m(i,j)| \quad (1)$$

Where B_n and B_m are the two frames of size N X M, and i and j denote pixel coordinates.

B. *Spatial feature extraction*

The spatial features extracted were the edge blocks, blurriness and the brightness between current and previous frames. Brightness is calculated as the modulus of difference between average brightness values of previous and current frames.

$$\mathrm{Br}_{n,}=\sum_{i=1}^{N}\sum_{j=1}^{M}\left|Br_{av(n)}(i,j)-Br_{\mathrm{av(n-1)}}(i,j)\right| \quad (2)$$

Where $Br_{av(n)}$ is the average brightness of *n*-th frame of size *N X M*, and *i* and *j* denote pixel coordinates.

The design of our content classification method is given in Fig. 3.

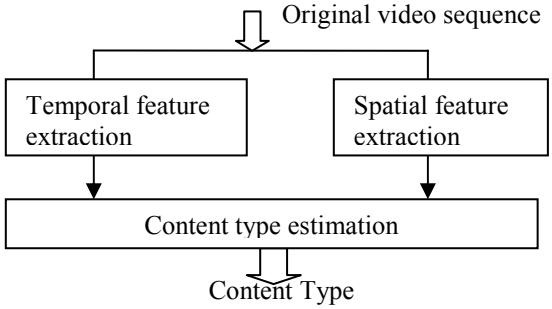

Fig. 3 Content classification design

C. *Cluster analysis*

For our data we calculate Euclidean distances in 10-dimensional space between the SAD, edge block, brightness and blurriness measurements and conduct hierarchical cluster analysis. Fig. 4 shows the obtained dendrogram (tree diagram) where the video sequences are grouped together on the basis of their mutual distances (nearest Euclid distance).

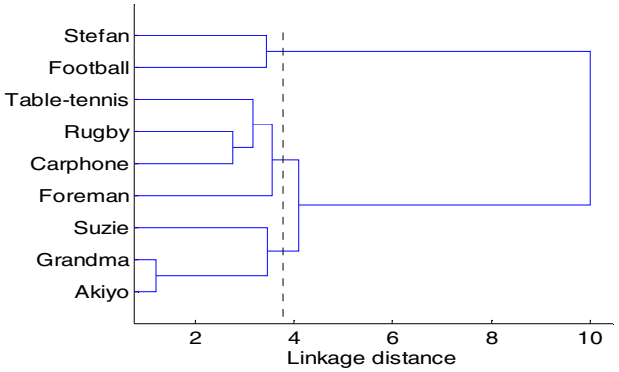

Fig. 4 Tree diagram based on cluster analysis

In this paper, we divided the test sequences at 38% from the maximum Euclid distance into three groups. (see the dotted line on Fig. 4) as the data contains a clear 'structure' in terms of clusters that are similar to each other at that point. Group 1 (sequences Grandma, Suzie and Akiyo) are classified as 'Slight Movement', Group 2 (sequences Carphone, Foreman, Table-tennis and Rugby) are classified as 'Gentle Walking' and Group3 (sequences Stefan and Football) are classified as 'Rapid Movement'. See Table III. Future work will concentrate on reducing the maximum Euclid distance and hence increase the content groups.

We found that the 'news' type of video clips were clustered in one group, however, the sports clips were put in two different categories i.e. clips of 'stefan' and 'football' were clustered together, whereas, 'rugby' and table-tennis' were clustered along with 'foreman' and 'carphone' which are both wide angle clips in which both the content and background are moving.

To further verify the content classification from the tree diagram obtained (Fig. 4) we carried out K-means cluster analysis in which the data(video clips) is partitioned into *k* mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. K-means computes cluster centroids differently for each measured distance, to minimize the sum with respect to the specified measure. We specified k to be three to define three distinct clusters. In Fig. 5 K-means cluster analysis is used to partition the data for the nine content types. The result set of three clusters are as compact and well-separated as possible giving very different means for each cluster. Cluster 1 in Fig. 5 is very compact for three video clips instead of four. The fourth clip of table-tennis can be within its own cluster and will be looked in much detail in future work. All results were obtained using MATLAB™ 2008 functions.

TABLE III
VIDEO CONTENT CLASSIFICATION

Content type	Content features	Video Clip
Slight Movement (SM)	A newscaster sitting in front of the screen reading news only by moving her lips and eyes	Grandma
		Suzie
		Akiyo
Gentle Walking (GW)	with a contiguous change of scene at the end – 'typical for video call'	Table-tennis
		Carphone
		Rugby
		Foreman
Rapid Movement (RM)	A professional wide angle sequence where the entire sequence is moving uniformly	Football
		Stefan

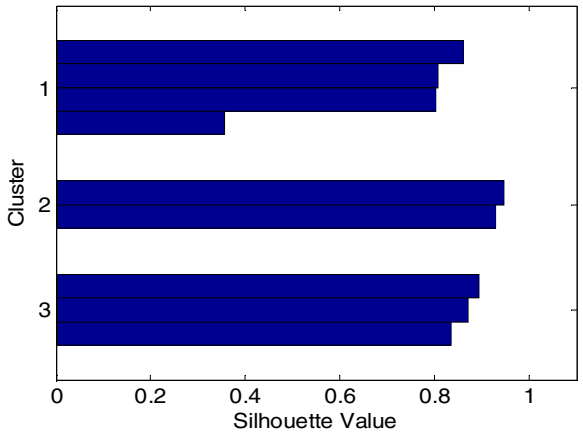

Fig. 5 K-means Cluster analysis

IV. Video Quality Prediction

The second step in this paper is to predict video quality based on the objective parameters of send bitrate, frame rate and packet error rate (see Table I) for the three content types of 'Slight movement', 'Gentle walking' and 'Rapid movement'. From the three content types classified in the previous section, we chose one video clip from each content type (see Table III) for testing purposes and a different video clip from the same content type for validation purposes. We chose video clips of 'Grandma', 'Foreman' and 'Stefan' from the three content types. For validation purposes we chose 'Suzie', 'Carphone' and 'Football'. Snapshots of the three video clips in the three content types used for validation are given in Fig. 6.

Fig. 6 Snapshots of three content types

Video quality prediction is carried out for three distinct content types from both network and application parameters for video streaming over wireless networks application as shown in Fig. 7. The application level parameters considered are Content Type (CT), Frame Rate (FR) and Send Bitrate (SBR). The network parameters are Packet Error Rate (PER).

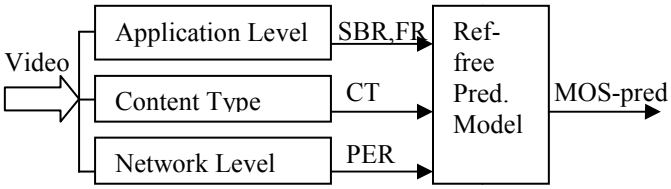

Fig. 7 Video quality metric prediction design

A. PCA Analysis

Principal Component Analysis (PCA) [12] reduces the dimensionality of the data while retaining as much information as possible. For this reason, PCA was carried out to determine the relationship between MOS and the objective video parameters of SBR, FR and PER. PCA involves calculating eigenvalues and their corresponding eigenvectors of the covariance or correlation matrix. Covariance matrix is used where the same data has the same set of variables and correlation matrix is used in the case where data has a different set of variables. In this paper, we used a covariance matrix because of the same data set.

TABLE IV
VAR. OF THE FIRST TWO COMPONENTS FOR ALL CONTENT TYPES

Sequence	Var. of PC1(%)	Var. of PC2(%)
Slight Movement	58	33
Gentle Walking	63	31
Rapid Movement	74	20

The PCA was carried out to verify the applicability of the objective parameters of SBR, FR and PER for metric design. The PCA was performed for the three content types of SM, GW and RM separately. The variance of the data for the three content types is given in Table IV. The first two components account for more than 90% of the variance and hence are sufficient for the modeling of the data. The PCA results are shown in Fig. 8.

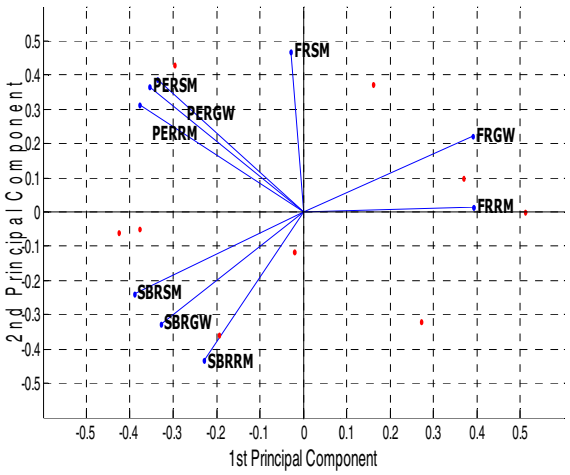

Fig. 8 PCA results for all content types

The PCA results from Fig. 8 show the influence of the chosen parameters (SBR, FR and PER) on our data set for the three content types of SM, GW and RM. In Fig.8 the horizontal axis represents the first principal component (PC1) and the vertical axis represents the second principal component (PC2). Each of the objective parameters (e.g. FRGW, etc) are represented by a vector.

B. MOS Prediction

The proposed low complexity metric is based on three objective parameters (SBR, FR and PER) for each content type as given by equation 3:

$$\text{MOS} = f(\text{SBR, FR, Content type, PER}) \quad (3)$$

We propose one common model for all content types given by equation (3). The prediction model for video quality evaluation is given by a rational model with a logarithmic function (see equation (4)).

$$MOS_v = \frac{a_1 + a_2 \text{FR} + a_3 \ln(\text{SBR})}{1 + a_4 \text{PER} + a_5 (\text{PER})^2} \quad (4)$$

The metric coefficients were obtained by a linear regression of the proposed model with our training set (MOS values obtained by objective evaluation given in Table I). The coefficients for all three content types are given in Table V.

TABLE V
COEFFICIENTS OF METRIC MODELS FOR ALL CONTENT TYPES

Coeff	SM	GW	RM
a1	4.5796	3.4757	3.0946
a2	-0.0065	0.0022	-0.0065
a3	0.0573	0.0407	0.1464
a4	2.2073	2.4984	10.0437
a5	7.1773	-3.7433	0.6865

The proposed metric has different coefficient values for the three different content types because spatial and temporal sequence characteristics of the sequences are significantly different. The model's prediction performance is given in terms of the correlation coefficient R^2 (indicates the goodness of fit) and the RMSE (Root Mean Squared Error) and is summarized in Table VI.

TABLE VI
METRIC PREDICTION PERFORMANCE BY CORRELATION COEFFICIENT AND RMSE

Content type	SM	GW	RM
Corr coef	79.9%	93.36%	91.7%
RMSE	0.2919	0.08146	0.2332

The performance of the video quality prediction obtained by our metric compared to the video quality data obtained objectively using NS2[15] for content type of 'gentle walking' is shown in Fig. 9. We achieved slightly better correlation for 'gentle walking' compared to the other two content types of 'slight movement' and 'rapid movement'. We also observed that video clips in 'rapid movement' are very sensitive to packet loss. The quality degrades rapidly compared to the other two categories as packet loss is introduced. Whereas, for 'slight movement' the video quality was still acceptable (MOS>3.5) for packet losses of up to 20%. Also compared to recent work published in [3],[6],[7] our results in terms of correlation coefficients and root mean squared error are comparable to theirs.

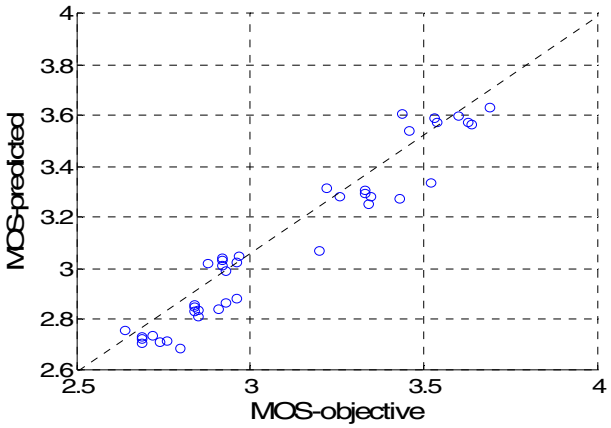

Fig. 9 Predicted vs. objective MOS results for 'GW'

V. CONCLUSION

In this paper we have proposed content based perceptual quality reference-free metric for the most frequent content types for wireless MPEG4 video streaming applications and investigated their performance.

We used cluster analysis to classify the most frequently used contents into three specific content types of 'slow movement', 'gentle walking', and 'rapid movement' based on the spatial and temporal feature extraction. The purpose of content clustering was to make new groups of video content with similar characteristics. The grouping could be used to apply priority control to content delivery, and hence optimize bandwidth allocation for specific content in content delivery networks. The automatic content classification can enable video quality prediction within a content type.

The proposed reference-free metric was validated with different video clips within the same content type with good prediction accuracy.

Future work will concentrate on collecting subjective MOS scores and extending our work to predicting and adapting the encoding parameters according to the content dynamics that satisfy a specific video quality level at a pre-encoding stage and hence propose an end-to-end perceived QoS framework based on both the application and network level parameters.

ACKNOWLEDGMENT

The work reported here is supported in part by the EU FP7 ADAMANTIUM project (contract No. 214751).

REFERENCES

[1] ITU-T. Rec P.800, Methods for subjective determination of transmission quality, 1996.
[2] K. Yamagishi and T. Hayashi, "Opinion model using psychological factors for interactive multimodal services", IEICE Trans. Communication, Vol.E89-B, No. 2, Feb. 2006.
[3] K. Yamagishi, T. Tominaga, T. Hayashi and A. Takashi, "Objective quality estimation model for videophone services", *NTT Technical Review*, Vol. 5, No. 6, June 2007.
[4] H. Koumaras, A. Kourtis, C. Lin and C. Shieh, "A theoretical framework for end-to-end video quality prediction of MPEG-based sequences", *Third international conference on Networking and Services*, 19-25 June 2007.
[5] P. Calyam, E. Ekicio, C. Lee, M. Haffner and N. Howes, "A gap-model based framework for online VVoIP QoE measurement", *Journal of Communications and Networks*, Vol. 9, No.4, Dec. 2007, pp. 446-56.
[6] A. Khan, L. Sun and E. Ifeachor, "An ANFIS-based hybrid video quality prediction model for video streaming over wireless networks", *Second Int. Conf. & Exhibition on Next Generation Mobile Applications, Services and Technologies(NGMAST)*, 16-19 Sept. , 2008, Cardiff, UK.
[7] M. Ries, O. Nemethova and M. Rupp, "Video quality estimation for mobile H.264/AVC video streaming", *Journal of Communications*, Vol. 3, No.1, Jan. 2008, pp. 41-50.
[8] L. Yu-xin, K. Ragip, B. Udit, "Video classification for video quality prediction", *Journal of Zhejiang University Science A,* 2006 7(5), pp 919-926.
[9] Y. Kato, A. Honda and K. Hakozaki, "An analysis of relationship between video contents and subjective video quality for internet broadcasting", *Proc. Of the 19th Int. Conf. On Advanced Information Networking and Applications (AINA),* 2005.
[10] Y. Suda, K. Yamori and Y. Tanaka, "Content clustering based on users' subjective evaluation", *6th Asia-Pacific symposium on Information and Telecommunication Technologies ASPITT*, 2005, Volume , Issue , 09-10 Nov. 2005 pp. 177 – 182.
[11] S. du Toit, A. Steyn and R. Stumpf, Cluster analysis", Handbook of graphical exploratory data analysis, ed. S.H.C. du Toit, pp.73-104, Springer-Verlag, New York, 1986.
[12] W. J. Krzanowski, "Principles of multivariate analysis", Clarendon press, Oxford, 1988.
[13] J. Wei, Video content classification based on 3-d Eigen analysis, *"IEEE transactions on image processing*", Vol. 14, No.5, May 2005.
[14] G. Gao, J, Jiang, J. Liang, S. Yang and Y. Qin, PCA-based approach for video scene change detection on compressed video, *Electronics letters*, Vol. 42, No.24, 23rd Nov. 2006.
[15] NS2, http://www.isi.edu/nsnam/ns/.
[16] J. Klaue, B. Tathke, and A. Wolisz, "Evalvid – A framework for video transmission and quality evaluation", *In Proc. Of the 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation,* Urbana, Illinois, USA, 2003, pp. 255-272.
[17] Ffmpeg, http://sourceforge.net/projects/ffmpeg
[18] J. Mitchell and W. Pennebaker, "MPEG Video: Compression Standard, Chapman and Hall, 1996, ISBN 0412087715.