Improving Mobile Video Telephony

Shraboni Jana, Eilwoo Baik, Amit Pande and Prasant Mohapatra
University of California, Davis, CA 95616
Email: {sjana, ebaik, pande, pmohapatra}@ucdavis.edu

***Abstract*—Video telephony is becoming popular over smartphones and tablets. Unlike the Desktop era, smartphone users are often 'mobile' and this impacts how the video is processed and transmitted over the network. The significant increase in the motion content in such videos change the composition of video frames. Coupled with wireless packet losses, it often leads to poor quality of video received by the end user. In this work, we propose RVD, a framework for Reliable Video Delivery in mobile telephony by accounting for video object motion comprising foreground end-user motion and background scene changes in the network transmission of video. Multilayer perceptron (MLP) based non-linear regression model is used to analyze the impact of redundancy on received video quality under network variations and different degrees of video motion. RVD achieves 17-25% bandwidth savings for a target video quality, and 50-56% quality improvement over video-oblivious approaches.**

I. Introduction

Mobile video telephony is one of the most sought-after real-time interactive application with its usage on the rise in both enterprise and consumer worlds [1]. Many applications such as Skype, Google Hangout, FaceTime, Fring, Tango, Vtok, ooVoo are becoming popular in the market.

The challenge for mobile video telephony applications is to provide reliable delivery of real-time videos on lossy links under strict delay constraints [2]. To combat such wireless link losses, FEC (Forward Error Correction) is the most efficient technique [3]. FEC is based on adding redundancy to the transmitted data at the expense of network bandwidth. Thus, it is important to apply only the required amount of redundancy to telephony videos.

The video content of video telephony applications have been conventionally perceived as confined to the user's facial expressions in static background, and are often modeled as 'head-and-shoulder' videos [4], [5]. Advent of smartphone has changed these notions about video telephony applications. Unlike the traditional desktop or video-conferencing scenarios, smartphone end-users have the flexibility to hold the end-device and move around. Thus, user movements can result in frequent background changes captured by the device camera.

The term 'mobility' has been often used in context of user mobility and its impact on network performance [6] and handover [7]. Those effects are not so grave in commonplace situations such as a person walking in a room or park while attending a video call. Thus, a more direct impact of user mobility in video telephony is in terms of 'motion' in the 'created' video content itself and its increased demands for network resources.

978-1-4799-4657-0/14/$31.00 © 2014 IEEE

Increased user motion may lead to frequent changes in video data. Failing to communicate these relevant video data to end user due to network losses can lead to long term impairment in visual quality [8]. The impact of user 'mobility' has not been studied properly, nor applied to improve reliability of video delivery services. In this paper, we propose RVD, an architecture for Reliable Video Delivery in real-time 'mobile' applications. To the best of our knowledge, this impact on video data and a strategy to provide reliable video delivery considering the impact has never been looked into before. The main contributions of our work are as follows:

- We characterize the impact of user/camera motion on generated video data in real-time telephony applications.
- RVD studies 'motion' in video content generated by mobile video telephony in terms of motion-vectors and scene-change information which is easily accessible in video codecs and needs no extra computations.
- RVD uses motion vector and scene change information, along with packet loss rate to derive a Multi-Layer Perceptron (MLP) based non-linear regression model which gives an accuracy of 95% in estimating the redundancy required to provide a desired video quality to end user.
- We implement RVD on test-beds and our results show -
  1) Performance improvement of received video quality by 50%-56% for 4%-10% network loss rates, in comparison to the approach oblivious of video data characteristics.
  2) RVD saves network bandwith utilization by 17%-25% with 4% -10% network loss rates, respectively, in comparison to the approach treating all video data having high motion vector values.

RVD provides insights to improve mobile video chat applications such as Skype. Skype currently uses FEC codes to add redundancy to video telephony [9]. However, in our prior work [10], we have shown how Skype videos with mobile clients have reduced video quality (higher blocking) in case of mobile scenarios as shown in Figure 1 [10]. RVD can be used in such applications to opportunistically decide the FEC redundancy levels and hence sending rate of the video-chat application depending on the network packet loss rate and user/camera motion to guarantee a high video quality to the end client.

The paper is organized as follows: In Section II, we discuss related works followed by telephony video classifications and understanding their distinguishing characteristics in Section III. Section IV discusses the impact of motion vector and scene change characteristics on the received video quality under variable loss-rates. In Section V, we provide detailed

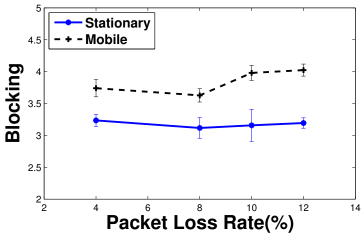


Fig. 1. Skype Video Quality under varying network conditions (Packet loss rates)

methodology to design RVD system including experiments and results followed by conclusion in Section VI.

II. Related Work

Real-time video data transmission over lossy wireless links is a challenging task [2]. Existing solutions use the following approaches to this problem: improving the efficiency and compression rates of codec or improving the reliability of network transmission or both.

Usach et. al [11] modulate the generated video information adapting frame-rate or adapting the bit-rate by varying frame quality for a given frame-rate. However, this requires frequent changes to the video codec parameters in real-time. The packet loss rate in wireless networks typically fluctuates depending on the interference, fading and noise. This may result in repeated changes in video coding parameters at the video codec, making it inefficient and/or too slow [12]. Moreover, such solutions are codec dependent.

The problem of reliable video delivery in network in presence of packet loss has been widely studied in the literature [13], [14], [15], [16], but these works do not differentiate real-time multimedia applications with video streaming applications. Due to the sensitivity to delay, re-transmissions and buffering techniques can not be applied to real-time multimedia communications. FEC (Forward Error Correction) codes are generally used for applications with strict delay constraints [3] to make them more resilient to packet loss. These codes introduce redundancy to existing bits of video stream.

Using inappropriate levels of FEC codes may lead to either lack of resiliency or else wastage of network resources. *How much FEC is required remains an open question?* Many FEC variants have been proposed [17], [18], [19], [20] to manage the amount of redundancy based on the networking conditions. The work in [17] focuses on adaptive FEC scheme for audio packets of internet telephony. In [18], the authors use block-based FEC codes which can be computationally intensive in real-time. The authors in [19], optimize FEC for layered video coding, but single-layer videos are more suitable for video telephony applications to reduce computational cost and latency [1].

The work in [20], [21] relies on frame type information - whether the packet represents I-frame, P-frame or B-frame. This analysis is valid for desktop-styled video telephony, but user mobility may sometimes reverse the relative importance of video frames. Higher information content in a frame should be the criterion for giving higher redundancy, not just classification into I, P or B categories. Based on the motion of video content, the video packets size vary proportionally.

To the best of our knowledge, none of the works address the issues of transmission of mobile video telephony and how we can leverage the video motion and scene changes information to reliably transmit video in lossy wireless links.

III. Telephony Video Characteristics

In this section, we study different types of videos generated during video telephony and their characteristics.

Video Classification

Real-time videos can be filmed with fixed or moving camera with objects in a camera having no motion to moderate motion. We classify the videos into four different groups, consistent with prior work [22].

1) Low-Mobility (LM) Videos - These videos contain end-users facial, head and shoulder movements. Background remains unchanged. These sequences are filmed with a fixed camera with relatively low motion on the scene. Video telephony has been conventionally considered to be LM videos [4], [5].
2) Medium-Mobility (MM) Videos - These videos are generated when end-users hold their smartphones and are walking during video conversation. The sequences have moderate scene changes and some pannings and other smooth camera movements. The videos also contain end-users facial, head and shoulder movements.
3) High-Mobility (HM) Videos - The background in video keeps changing continously though foreground is focused on end-users head, face and shoulder. These sequences have high motion scenes filmed with a camera which is in continuous movement.
4) Mixed (MX) Videos - These videos which have sequences that have combinations of all types of situations mentioned above.

The set of all videos is represented as V_{type}, where, $V_{type} = \{LM, MM, HM, MX\}$. For our work we recorded 20 video clips for each video type using Galaxy Nexus S smartphone. The videos are encoded to H.264/AVC videos with 352x258 resolution at 29.97 frames per second (fps) using FFMPEG [23]. The GOP size was chosen to be 250 to enable high compression rates. We validated our results for this section at high resolution (720p) and small GOP sizes (30) and observed consistent results.

We use Peak Signal-to-Noise Ratio (PSNR) metric to determine the quality of received video. PSNR is the simplest and most widely used video quality evaluation methodology, defined as ratio between the maximum possible power of video signal (square of maximum pixel value, 255) and the power of corrupting noise that affects the fidelity of its representation. It is measured in dBs. In our case, the corrupting noise

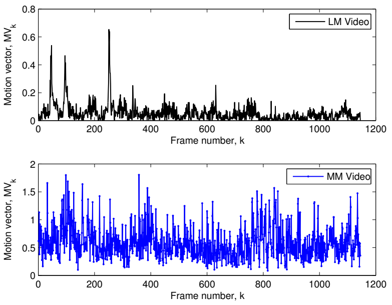


Fig. 2. Plot of MV_k per frame f_k for LM and MM videos.

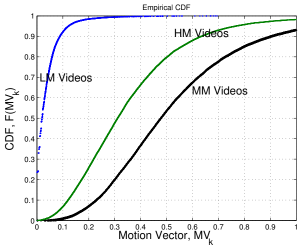


Fig. 3. CDF of motion Vectors MV_k per frame f_k for LM, MM and HM videos.

originates from the packet losses in network transmission. Mathematically, PSNR is calculated between two video frames

$$PSNR = 10\log_{10}\frac{3MN \times 255^2}{\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{d=1}^{3}(S(i,j,k)-R(i,j,k))^2}$$

where $S(.)$ and $R(.)$ are the source and received video frames respectively. MxN, is the resolution in two dimensions and 3 is color depth.

Understanding Mobile Videos

We, henceforth, study the properties of LM, MM and HM videos in terms of their average motion vectors and scene changes. Though we discuss and demonstrate the efficacy of our approach using the ITU-T H.264 video codec, our method is valid for most video codecs.

First, let us review how video frames are coded and compressed. Recent video coding standards use intra and inter-frame predictions to exploit the temporal and spatial correlations found in natural image sequences for bit-rate reduction. Intra-frames (I-frame or reference frame) are coded independently using only information present in the frame/picture itself where as inter-frames (P or B-frames) are predicted from neighboring intra and inter-frames. These frames are then grouped into g number of frames known as GOP (Group of Pictures) with an intra-frame followed by $g-1$ inter-frames [24]. To maintain low-latency, baseline profile (I and P-frames only) is considered for video telephony applications [25].

A frame in a video is divided into slices. A slice in turn is divided into macroblocks. I-slice contains intra-coded macroblocks only. In P-frame slices, each macroblock may be either coded using inter-prediction or intra-prediction or skipped. The macroblocks may be split into partitions of size - 16x16, 16x8, 8x16 or 8x8 pixels. These partitions can be divided into sub-partitions, for example, an 8x8 partitions can have sub-partitions as- 8x8, 8x4, 4x8 and 4x4. The choice of partitions, sub-partitions, inter and intra-prediction for macroblocks are determined by rate-distortion optimization scheme used by video codec.

Inter-coded macroblocks in a P-slice are predicted from a number of previously coded frames using motion compensation. A single motion vector is required for each of such partition or sub-partition. The absolute motion vector, amv, for the i^{th} macroblock in k^{th} frame is evaluated as [24] follows -

$$amv_i = \sum_{j=0}^{15} |mv^x(i,j)| + |mv^y(i,j)| \tag{1}$$

where $mv^x(i,j)$ and $mv^y(i,j)$ are the displacements in x- and y- direction of j^{th} 4×4 sub-block in the i^{th} macroblock. Hence, the sum of absolute motion vectors (MV_k) of the k^{th} frame with N inter-coded macroblocks is

$$MV_k = \sum_{i=0}^{N-1} amv_i \tag{2}$$

This is indicative of frame motion quantified and stored in the form of motion vectors. Figure 2 plots motion vector values of each video frame for LM and MM videos. The MV values clearly indicate that MV values are higher for MM compared to LM. Figure 3 shows the CDF (Cummulative Distributed Function) of motion vectors for LM, MM and HM videos. We find that the motion vector values are non-linearly related to video type. Hence, the relationship between v, where, $v=\{x|x\in V_{type}\}$ and average MV values, $\widehat{MV}$ is derived using non-linear decision tree (REPTree [26]) algorithm for approximately, 72,000 video frames as following

$$\widehat{MV} < 0.19 \quad \forall \quad v = LM \tag{3}$$
$$0.19 \leq \widehat{MV} < 0.46 \quad \forall \quad v = HM \tag{4}$$
$$\widehat{MV} \geq 0.46 \quad \forall \quad v = MM \tag{5}$$

The model has both Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as 0. It is evident that the motion vectors of MM videos have considerably higher values than LM videos whereas the motion vectors of HM videos have lower values than MM videos, the reason being discussed later in this section.

Intra-coded P-slice macroblocks are encoded in spatial domain using blocks of pixels that are already encoded within the current frame and requires much larger bits than inter-coded predicted macroblocks. Thus, for a P-frame, if number of intra-coded macroblocks are more, then the scene change has been detected in the video. Scene change in the video is an active research area and many approaches have been laid down in the literature and is not a part of our study. To keep in line with our codec-independent approach, we use Sum of Absolute Differences (SAD) metric [24] to detect the scene

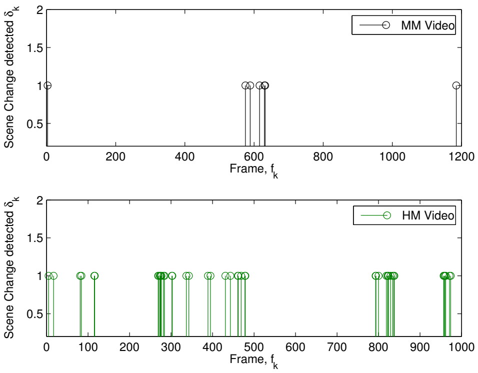


Fig. 4. Plot of δ_k per frame f_k in MM and HM videos. LM videos (not shown here) have $\delta_k = 0 \; \forall \; f_k$.

changes.

$$SAD_k = \sum_{i=1}^{4}\sum_{j=1}^{4} |a_{ij} - \tilde{a_{ij}}| \tag{6}$$

SAD is one of the simplest similarity measures which is calculated by subtracting pixels $\tilde{a_{ij}}$ of frame f_{k-1} (reference-frame) and the pixels a_{ij} of the target frame f_k followed by the aggregation of absolute differences. High SAD values are generated on scene changes, thereby, making such changes detectable. If $SAD_k > \eta$ where η is pre-defined threshold, then we can conclude that the scene has changed i.e. $\delta_k = 1$ and 0 otherwise. Using empirical studies, we find $\eta = 60$ suitable for detecting scene changes.

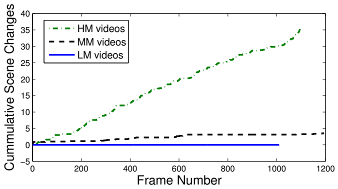


Fig. 5. Cumulative plot showing cumulative SCD events as a function of frame number.

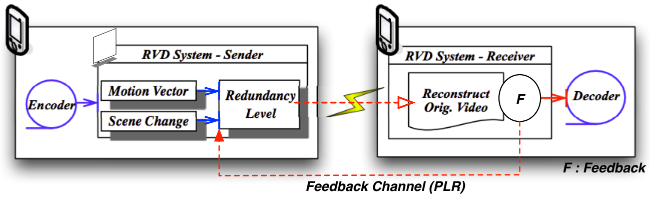


Fig. 6. Overall architecture for proposed RVD system.

TABLE I
SYSTEM PARAMETERS

Symbols	Descriptions
k	Frame Number.
f_k	k^{th} video-frame.
g	GOP size.
MV_k	Motion vector of k^{th} frame.
δ_k	Scene Change detected in frame f_k.
plr	Packet loss rate.
R	Redundancy Level $\forall k$.

Figure 4 plots the scene change event in an arbitrary video trace of MM and HM videos. Figure 5 shows average cumulative scene changes, $\widehat{\delta} = \frac{\sum_{j=1}^{N}\sum_{i=1}^{k}\delta_i}{N}$ detected for k frames for each type of videos having N test videos. LM videos detect no scene changes consistently for all test videos. We find that scene changes detected in the HM videos is higher compared to MM videos where as motion vectors have higher values in comparison to HM videos (Figure 3). The reason being the HM videos have more frequent background changes. Hence, with a scene change, most macroblocks are intra-coded and there are very few or even no inter prediction motion vectors. On the other side, the more macroblocks are inter-coded, the more likely the current frame is correlated with the previous frame and has no scene change thus resulting in higher values of motion vectors for MM videos. This inference is counter-intuitive, as one would expect linearly higher values of motion vectors in higher user mobility videos.

Inferences

1) Motion Vector (MV) values are higher in MM videos. The average motion vectors for LM, MM and HM videos are 0.0360, 0.5321 and 0.3552 respectively (Figure 3).
2) Scene Change Detection (SCD) is higher in HM videos. The average scene change detected for LM, MM and HM videos are approximately 0, 5 and 30 respectively (Figure 4) for 1000 video frames.

IV. RVD SYSTEM

Figure 6 gives block diagram of proposed system. The video encoder in Figure 6 encapsulates the frame slices into Network Abstraction Layer (NAL) units. These NAL units are suitable for transmission over packet networks [3] using RTP (Real-time Protocol). RTP/UDP/IP is the dominant standard used for video telephony applications [1]. RVD sender system along with NAL units receives the following additional information from video encoder - (1) motion vector information per frame and (2) scene change detected per frame. The macroblock motion vectors are in any case generated as a part of the video compression at the video encoder. SAD (Sum of Absolute Differences) used for scene change detection is a metric available in most of the codecs and is the simplest and less expensive metric [27]. Hence obtaining information of these two metrics from video encoder does not require any changes in the video codec itself. RVD thus receives MV_k, δ_k from video encoder and network loss rate, plr from receiver via RTCP (Real-time Control Protocol). We assume plr remains constant during the transmission of all packets corresponding to video frame f_k.

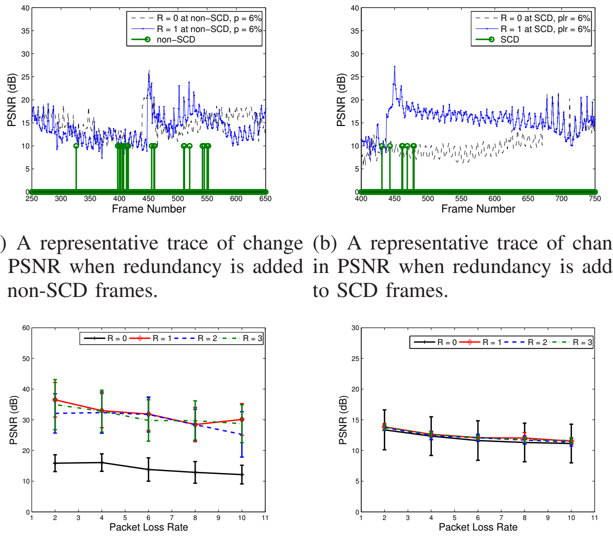
We use extensive simulations to study the video quality at the receiver for different video types by varying network loss rates and applying different degrees of redundancy. We design RVD algorithm based on simulations data and perform experiments on real test-bed discussed in later sections.

The MV and SCD information are obtained using FFMPEG. The simulations are carried using NS3 QoE Monitor [28]. NS3 QoE Monitor provides the network interface to examine the impact of networking conditions on received video quality. The point-to-point wireless link is set to 20Mbps with 2ms delay and maximum transmission unit as 1400 bytes. We assume that the packet loss probability of the wireless link is time-varying with error link model having uniform distribution. Such feedback about wireless link can be made available to the sender from receiver by using RTCP [29].

We, hereby, examine the impact of MV and SCD on received video quality for LM, MM and HM videos.

A. Impact of Scene Changes

The HM telephony videos have on an average 30 scene change detected (Figure 4) for approximately 40 seconds of captured videos. We examine the impact of ensuring the transmission of scene change packets on video delivery. Figure 7, hereby, shows the performance of adding redundancy to SCD frames and non-SCD frames in transmitted video. In all figures, PSNR imply average PSNR across all frames of received video in dB. When non-SCD frames are given redundancy (R), PSNR of such frames improve marginally, but the video quality of the subsequent frames have no improvements as shown in Figure 7(a). The lower the PSNR values, the more is the video quality degradation. The maximum PSNR of received videos is limited to 99dB [28] for plotting purposes.



(a) A representative trace of change in PSNR when redundancy is added to non-SCD frames.

(b) A representative trace of change in PSNR when redundancy is added to SCD frames.

(c) Impact of different redundancy levels on PSNR for HM videos.

(d) Impact of different redundancy levels on PSNR for MM videos.

Fig. 7. Mitigating the impact of scene changes in video quality by added redundancy.

Figure 7(b) shows quality improvement with the addition of redundant packets for SCD frames ($k \in \{431, 443, 461, 462, 469, 478, 479\}$) in a sample HM video. PSNR is improved by 10dB for subsequent video frames. This increase in PSNR ceases to exist around $k = 750$ as the videos considered in our simulations have maximum GOP size, $g = 250$. If SCD is detected in $\tilde{g}$ position, the reliable delivery of SCD frame packets likely improves the performance of $g - \tilde{g}$ frames till next GOP.

With addition of redundant packets for SCD frames (Figure 7(c)), the PSNR on an average increases for HM videos. The errorbars in all figures henceforth represent two standard deviation of the mean PSNR. However, the increase in redundancy levels (R = 2) does not further improve the video quality.

The distribution of SCD frames in a video is bursty by nature (Figure 4). The average number of SCD frames obtained from $\approx$ 40 second videos is 30 which corresponds to 200 packets on average, where as, average number of packets generated by these videos are $43,880$. SCD frames, thus comprises only 0.46% packets of total packets transmitted. Providing reliable delivery to only 0.46% packets with increased redundancy, intuitively, is not going to effect the global video quality beyond certain level and hence explains the result shown in Figure 7(c).

MM videos have overall no impact on received PSNR for redundancy addition on scene changes (Figure 7(d)). This can be attributed to the fact, MM videos have very few scene changes. Hence, the overall impact on video quality is negligible.

Inferences

1) Adding redundancy to SCD frames does improve quality of subsequent frames till next GOP.
2) SCD based redundancy addition is suitable for HM videos.
3) The amount of redundancy level added, has no impact on the video quality i.e. the performance saturates for $R \geq 1$.

The SCD results are based on 2840 video transmission sessions under variable conditions. As we mentioned earlier, efficient SCD algorithm is an active area of research and is out of the scope of our work. In RVD, we take advantage of existing SCD algorithms for providing content-aware reliable delivery to encounter network losses for mobile video telephony.

B. Impact of Motion Vector

The performances of LM, MM and HM videos in terms of PSNR is shown in Figure 8(a). It is evident, that the PSNR (in dB) degradation for MM videos is higher in comparison to other two video types for same packet-loss rates. PSNR degradation of LM videos is the least. Intuitively, loss of video data with less variations can be recovered by the decoder based on past video frames. We discussed in Section III, that MM videos have higher motion vector values compared to LM and HM videos. Hence, from Figure 8(a), we observe there is a positive correlation between PSNR degradation and motion vectors.

Figures 8(b), 8(c) and 8(d) show performances of LM, MM and HM videos for R = 1, 2 and 3. The redundancy level of more than 3 is not applied since the average PSNR for all video types is already more than 80dB. The performance improvement for different degrees of redundancy differs based on the video type. MM videos have least improvement in quality in comparison to LM and MM videos. The results of simulations are further summarized in Figures 8(e) and 8(f). The increase in video quality of LM videos relative to MM videos, referred as LM-MM for $R = 1$ is as high as 70%. However, as R is increased, the disparity in video quality for LM-MM decreases significantly for lower loss rates (<=6%). The MM videos having highest motion vector values also performs poorly in comparison to HM videos as shown in Figure 8(f). For higher loss rates (plr), the percentage PSNR improvement of LM and MM videos over MM is higher.

Inferences

1) MV is positively correlated to the received video quality degradation.
2) Variable degrees of redundancy is required depending on video motion for maintaining video quality at the receiver.

V. RVD Methodology

Our primary aim is to build a video delivery system for video telephony applications taking into consideration video content and network resources. The objective of the RVD system is to minimize the usage of network resource and maximize video quality at the receiver for video telephony applications. In total, 10,778 video transmission sessions are used to identify and examine features per video frame of RVD system. In this section, we first briefly introduce the regression techniques used to analyze and model the system followed by the impact of system features. The performances of these models are measured in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

A. Prediction Model

Since the simulation results indicate non-linearity between RVD system parameters, therefore, we apply Multiple Layer Perceptron (MLP), a non-parametric, non-linear and data driven machine learning approach to model the weights of RVD system. MLP is an example of an Artificial Neural Network represented as a finite directed acrylic graph comprising sets of input (Y), output (X) and hidden nodes. The weights of hidden nodes organized in layers is determined using supervised learning (also referred to as training). The process involves initializing the hidden nodes weight matrix, W_{xy}, between input and output nodes.

$$Y = f(X, W_{x,y}) \tag{7}$$

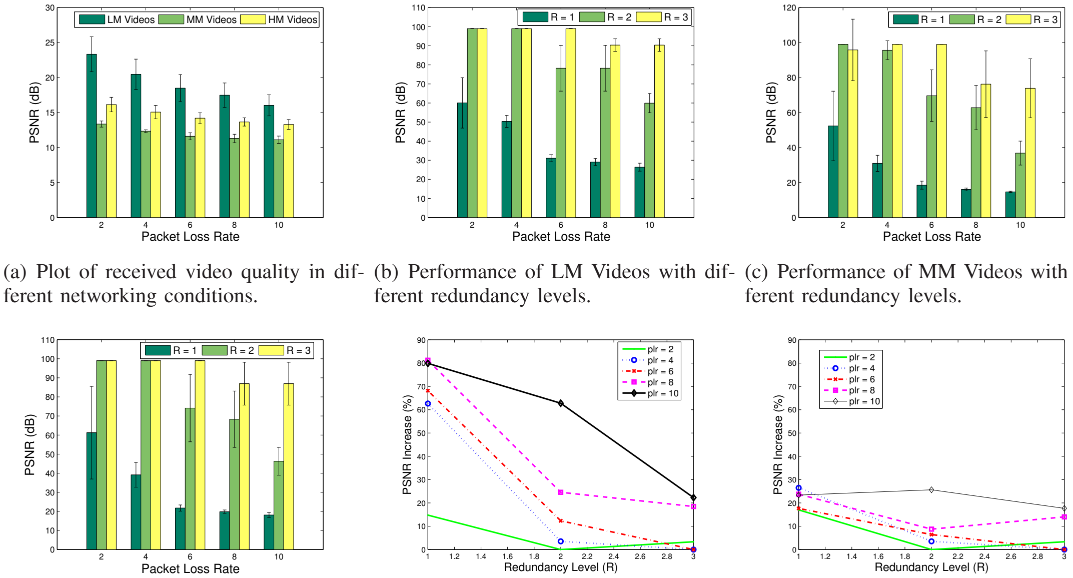
The weights are changed during training in order to minimize the error between the estimated MLP output and the correct value of output. The weights at each node is computed as -

$$X \rightarrow f_{log}(W_0 + < W, X >), \tag{8}$$

$$f_{log}(z) = \frac{1}{1 + e^{-z}} \tag{9}$$

where, f_{log} is known as sigmoid function. MLP uses gradient descent techniques for error minimization. Further MLP details can be referred in [26].

Using WEKA toolbox, we divide the data for the model into $n = 10$ folds, where, $n - 1$ folds are for supervised learning and one fold is used to test the model for errors. The errors obtained in a fold is added to the weights of nodes of next fold in the training set. 10-fold cross validation is used to build a robust model.



(a) Plot of received video quality in different networking conditions.

(b) Performance of LM Videos with different redundancy levels.

(c) Performance of MM Videos with different redundancy levels.

(d) Performance of HM Videos with different redundancy levels.

(e) Increase in PSNR(%) for LM videos with respect to MM videos

(f) Increase in PSNR(%) for HM videos with respect to MM videos

Fig. 8. Impact of added redundancy on videos with different motion.

B. System Features

System features of feature vectors are the inputs we give to MLP. We identify several input factors:

Target PSNR (Q_t): Target PSNR is required video PSNR at the receiver to maintain perceptual video quality at the end-user. The importance and computation of PSNR has been elaborated in Section III.

Target sending rate (S_t): The data stream generated by LM, MM and HM videos are between 1Mbps-2Mbps. The sending rate at the transmitter, intuitively, increases n-times if there are n redundant transmissions for each packet. Hence, S_t can impose major constraint for RVD system.

Target video frame jitter (J_t): We consider worst case scenario with $plr = 10\%$ to examine impact of added redundancy on jitter. Figure 9 shows impact of MV based redundancy on jitter experienced by video frames. The video frame jitter is computed as the difference in RTP time-stamps of first packet of a video frame to the last packet of the same video frame. As per [30], in high-quality videoconferencing all packets of each video frame are expected within 33 ms.

The impact of video frame jitter is observed in Figure 9. Since the GOP size is 250 for the test videos, the video jitter for I-frames spikes to 30-35 ms for $R = 3$ as I-frames have larger sizes. Thus for video data, even if packet level jitter is within limits, video frame jitter may not be within limits which is more relevant for video applications. Hence, target video frame jitter, J_t should be taken into consideration for evaluation of R.

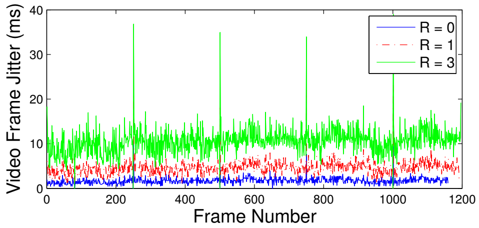


Fig. 9. Impact of redundancy addition on video frame jitter at packet loss rate = 10%.

SCD frame (δ_k): Since SCD-based redundancy addition applies to selective frames, the sending rate and jitter per frame packets are increased marginally. However, the performance gains for such redundancy is not more than 10dB (Figure 7(c)). Hence, based on sending rate, jitter constraints per video frame and target PSNR, RVD can switch between MV and SCD based redundancy for HM videos, whereas, redundancy for LM and MM videos are based on MV.

Motion: The impact of motion vector ($\widehat{MV}_k$) on video quality under variable network conditions is observed in Figure 8. The received video quality of videos having highest values of motion vectors deteriorates more in comparison to the videos having smaller motion vector values under same networking conditions.

In mixed videos, the user/camera movements may lead to one or more type of video sequences. We, therefore define $\widehat{MV}_k$, the value of average motion vector till k^{th} video frame. It is computed as $\widehat{MV}_k = \frac{\sum_{k=K-\omega}^{K} MV_k}{K}$, where, ω is a fixed window to be observed for determining average motion vector of the video till k^{th} frame.

Algorithm 1 RVD

```
Input $Q_t$, $S_t$ and $J_t$
for each frame $f_k$
Objective - Minimize $R_k$
Obtain $MV_k$, $v_k$, $\delta_k$ and $plr_k$
switch($v_k$)
case(HM)
if ($Q_t \leq Q_{scd}$) then
    Evaluate $R_k = scdlevel(\delta_k)$
else
    Evaluate $R_k = mvlevel(\widehat{MV}_k, plr_k, Q_t, S_t, J_t)$
end if
otherwise
if $v_k == v_{k-1}$ then
    $R_k = R_{k-1}$
else
    Evaluate $R_k = mvlevel(\widehat{MV}_k, plr_k, Q_t, S_t, J_t)$
end if
End
```

Packet loss rate (plr_k): We assume packet loss experienced per frame packets (plr_k) during transmission remains constant.

C. RVD Algorithm

Algorithm 1 discusses our proposed RVD algorithm. Since telephony videos are generally mixed videos - a combination of one or more video types (LM, MM and HM videos), it becomes essential to compute our algorithm at each video frame level. The objective of RVD system is to thus, minimize redundancy level required per frame, R_k for received PSNR $\geq Q_t$, transmission rate $\leq S_t$ and video frame jitter $\leq J_t$. For HM, RVD selects SCD based redundancy ($scdlevel(.)$) or MV based redundancy ($mvlevel(.)$) depending on the system constraints. Q_{scd} is the maximum PSNR achieved applying $scdlevel(.)$. Our simulation data shows $scdlevel(\delta_k) = \delta_k$ where as $mvlevel(\widehat{MV}_k, plr_k, Q_t, S_t, J_t)$ is modeled as MLP described above. v_k is the video type (v) in which k^{th} frame is classified and is a function of $\widehat{MV}_k$ (Equations 3 - 5).

D. Experiments

In this section, we discuss RVD implementation and its results. Mixed videos, captured using Nexus S, are used in implementation phase to test the efficacy of proposed RVD. We configure a single hop wireless testbed between a sender and a receiver. The configuration of WLAN in the testbed is IEEE 802.11n and the average minimum endto-end delay is 0.504 milli-seconds. FFMPEG is used for video encoding and decoding. The real-time video transmission is established by using VLC player. A FreeBSD iMAC server (3.2GHz Intel Core i5 with 8GB) was used for a sender and FreeBSD-based

TABLE II
PERFORMANCE OF SYSTEM FEATURES.

System Features	MAE	RMSE
$\{\widetilde{MV}_k, Q_t, J_t, S_t, plr\}$	0.1153	0.1692
$\{v, Q_t, J_t, S_t, plr\}$	0.1284	0.187
$\{Q_t, plr\}$	0.3702	0.4679

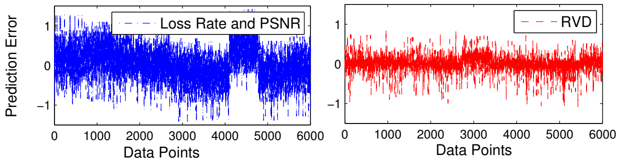


Fig. 10. Prediction error (of R) without and with accounting for video motion.

system (2.7GHz Intel Core i3 with 4GB) for a receiver. During the video transmission session, we use IPFireWall tool in order to introduce desired packet loss rates in the link.

Table II shows the MAE and RMSE values for the MLP model considering different combinations of system features. If we consider MLP with input $X = \{plr, PSNR\}$ and $Y = R$, then the trained model has a high prediction error in comparison to RVD as shown in Figure 10. The error for this model frequently fluctuates between ± 1. Since R can take values from $\{0, 1, 2, ...\}$, such a model will lead to inaccurate R for the system. Table II, also gives MAE and RMSE values for different combinations of inputs. We can see how MAE increases significantly (3X) if we only account for channel loss rate and not other input variables. The performance of the models considering system features - MV or video type do not differ in terms of MAE and RMSE. Thus, we can infer that MV values can be replaced by video types as a system feature to build MLP model.

Instead of implementing MLP on a smartphone or laptop, we port input-output relationships into a look-up based on Algorithm 1. A condensed version of look-up table is shown in Table III. $Q_t = 31$ dB was chosen to indicate modest video quality [31] while $Q_t = 50$ dB indicate high visual fidelity of received videos. It is evident from Table III that videos with higher values of motion vectors may require higher redundancies. For $Q_t = 31$ dB, LM videos require no redundancy for loss-rates 2% and 4%, whereas, MM and HM videos require $R = 1$ to achieve the target PSNR of 31 dB. The required redundancy levels are also observed different for 8% and 10% loss-rates. HM videos, having lower values of motion vectors in comparison to MM videos at $plr = 8\%$ achieves 31 dB with a reduced redundancy level than MM videos. With required $Q_t = 50$ dB (high fidelity videos) and $plr = 4\%$, the redundancy level of MM is more than other two video types whereas, LM videos require lower redundancy level in comparison to MM and HM videos.

Next, we use these values to stream videos over network. We consider four cases : 1) No-R implies no redundancy is applied 2) RVD 3) Min-R implies minimum redundancy applied i.e. entire video data is classified as LM and 4) Max-R implies maximum redundancy applied considering entire video as MM. Figure 12(a) shows RVD and Max-R, both consistently perform better than target PSNR, Q_t, where as, bandwidth consumed by Max-R is greater than RVD as shown in Figure 12(b). Performance improvement of received video quality by 50%-56% for 4%-10% network loss rates, in comparison to the approach oblivious of video data characteristics (No-R). RVD saves network bandwith utilization by 17%-25% with 4% -10% network loss rates, respectively, in comparison to the approach treating all video data having high motion vector values (Max-R).

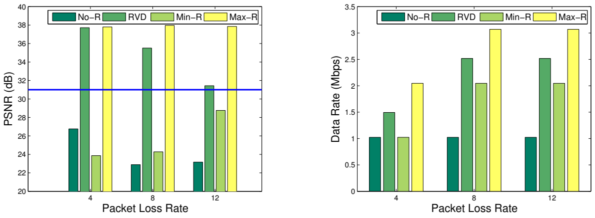
TABLE III
REDUNDANCY VALUES

Target PSNR = 31dB

$plr(\%)$	2	4	6	8	10
LM	0	0	1	1	1
MM	1	1	1	2	2
HM	1	1	1	1	2

Target PSNR = 50dB

$plr(\%)$	2	4	6	8	10
LM	1	1	2	2	2
MM	1	2	2	3	3
HM	1	1	2	3	3



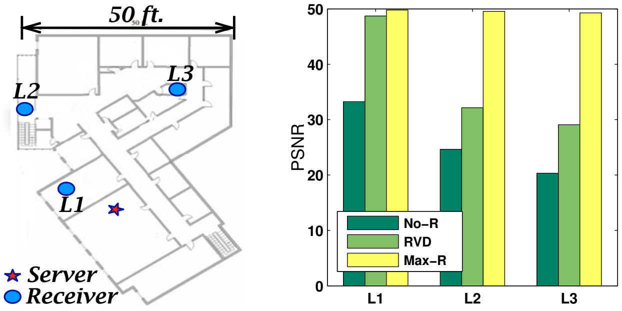
(a) Video Quality at receiver (b) Bandwidth overhead at sender

Fig. 11. Experiment results using mixed videos.

Next, we study how the performance varies across different locations on the second floor of a two-storey building at UC Davis (see Figure 12). The location of transmitter and receiver not only impact how signal propagates over the time but more importantly affect the selective fading at spectral domain. For desired PSNR threshold of 31 dB (Q_t) we observe that RVD is able to achieve this performance in all scenarios while No-R leads to poor performance in L2 and L3. Max-R is able to achieve high quality (50 dB) but leads to over twice the bandwidth requirements ($2X$) of RVD. The results are averaged over 12 video sessions for each location and scheme. L1 has low losses, hence No-R scheme also approaches 31 dB performance, while RVD quality is close to 50 dB. In these results, we have limited the maximum PSNR to 50 dB ($Q_t > 40$ dB represents excellent quality and any distortion is invisible to human eye [31]).

VI. CONCLUSION

In this work, we propose a scheme for intelligent video transmission by adjusting redundancy levels based on video motion. We focus on two main attributes of video information -



(a) Map

(b) Video Quality

Fig. 12. RVD performance in different locations ($Q_t = 31$ dB)

scene changes and motion vectors. The videos with high values of average motion vectors have higher degradation in video quality and require additional redundant packet transmissions. Scene change frames need additional redundancy for high quality.

With extensive simulations and experiments, we conclude that video content plays a vital role in improving video quality at the receiver. Such insights can be exploited by the video telephony application providers to fine-tune the video transmissions. RVD gives better trade-off between bandwidth usage and received video quality in comparison to approaches which are oblivious of video content. We used standard FEC codes to demonstrate the impact of video motion. However, advanced redundancy mechanisms such as Raptor Codes or Selective Re-transmission can also be used to further reduce the required data redundancy.

REFERENCES

[1] S. Jana, A. Pande, A. Chan, and P. Mohapatra, "Mobile video chat: Issues and challenges," *IEEE Communications Magazine*, 2013.

[2] M. van der Schaar and N. Sai Shankar, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, 2005.

[3] S. Wenger, M. M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, "RTP Payload Format for H.264 Video," Internet Engineering Task Force, Tech. Rep., feb 2005.

[4] D. Chai and K. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Transactions on Circuits and Systems for Video Technology,*, vol. 9, no. 4, pp. 551–564, 1999.

[5] V. Bruce, "The role of the face in communication: Implications for videophone design," *Interacting with computers*, vol. 8, no. 2, pp. 166–176, 1996.

[6] M. Vutukuru, H. Balakrishnan, and K. Jamieson, "Cross-layer wireless bit rate adaptation," in *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4. ACM, 2009, pp. 3–14.

[7] S. Fernandes and A. Karmouch, "Vertical mobility management architectures in wireless networks: A comprehensive survey and future directions," *IEEE Communications Surveys & Tutorials,*, vol. 14, no. 1, pp. 45–63, 2012.

[8] J. Greengrass, J. Evans, and A. Begen, "Not all packets are equal, part 2: The impact of network packet loss on video quality," *IEEE Internet Computing*, vol. 13, no. 2, pp. 74–82, 2009.

[9] T.-Y. Huang, P. Huang, K.-T. Chen, and P.-J. Wang, "Could skype be more satisfying? a QoE-centric study of the FEC mechanism in an internet-scale VoIP system," *Netwrk. Mag. of Global Internetwkg.*, vol. 24, no. 2, pp. 42–48, Mar. 2010.

[10] S. Jana, A. Pande, A. Chan, and P. Mohapatra, "Network characterizationd and perceptual quality of skype mobile videos," in *Proc. of ICCCN*, 2013.

[11] P. Usach, J. Sastre, and J. Lopez, "Variable frame rate and gop size h.264 rate control for mobile communications," in *IEEE International Conference on Multimedia and Expo*, 2009.

[12] S. Wenger, B. Burman, L. Hamm, and Ericsson, "Codec operation point rtcp extension," Internet Engineering Task Force, Tech. Rep., mar 2012.

[13] R. Han and D. Messerschmitt, "A progressively reliable transport protocol for interactive wireless multimedia," *Multimedia Systems*, vol. 7, no. 2, pp. 141–156, Mar. 1999.

[14] M.-H. Lu, P. Steenkiste, and T. Chen, "Robust wireless video streaming using hybrid spatial/temporal retransmission," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 476–487, 2010.

[15] J. Apostolopoulos, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," in *Visual Communications and Image Processing (VCIP)*, 2001, pp. 392–409.

[16] N. Feamster and H. Balakrishnan, "Packet loss recovery for streaming video," in *In 12th International Packet Video Workshop*, 2002.

[17] J.-C. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive fec-based error control for internet telephony," in *Proc. of INFOCOM*, 1999.

[18] J. Xiao, T. Tillo, C. Lin, and Y. Zhao, "Real-time forward error correction for video transmission," in *Proceedings of IEEE VCIP*, 2011, pp. 1–4.

[19] C. Hellge, D. Gomez-Barquero, T. Schierl, and T. Wiegand, "Layer-aware forward error correction for mobile broadcast of layered media," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 551–562, 2011.

[20] Y. C. Chang, S.-W. Lee, and R. Komyia, "A fast forward error correction allocation algorithm for unequal error protection of video transmission over wireless channels," *IEEE Transactions on Consumer Electronics,*, vol. 54, no. 3, pp. 1066–1073, 2008.

[21] C.-H. Lin, Y.-C. Wang, C.-K. Shieh, and W.-S. Hwang, "An unequal error protection mechanism for video streaming over IEEE 802.11e WLANs," *Computer Networks*, vol. 56, no. 11, pp. 2590–2599, Jul. 2012.

[22] J. Sastre, P. Usach, A. Moya, V. Naranjo, and J. Lopez, "Shot detection method for low bit-rate H.264 video coding," in *Proc. of EUSIPCO*, 2006.

[23] FFMPEG. [Online]. Available: http://ffmpeg.org/

[24] I. E. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. Wiley. com, 2004.

[25] J. Wang, "ChitChat: Making Video Chat Robust to Packet Loss," Tech. Rep., July 2010.

[26] I. H. Witten and E. Frank, *Data Mining, A Practical Maching Learning Tools and Techniques*. Elsevier, 2005.

[27] A. M. Tourapis, "Enhanced predictive zonal search for single and multiple frame motion estimation," in *Proc. of SPIE*, 2002.

[28] "A tool for multimedia quality assessment in NS3: QoE Monitor," *Simulation Modelling Practice and Theory*, vol. 32, no. 0, pp. 30 – 41, 2013.

[29] J. Ott, S. Wenger, and et. al, "Extended rtp profile for rtcp-based feedback (rtp-avpf) internet draft," Tech. Rep., 2002.

[30] R. S. T. Szigeti, K. McMenamy and A. Glowaski, *Cisco TelePresence Fundamentals*. ciscopress, 2009.

[31] J. Klaue, B. Rathke, and A. Wolisz, "Evalvid - a framework for video transmission and quality evaluation," in *Proc. of International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, 2003, pp. 255–272.