Quality Estimation for Scalable Video Codec

Presented by
Ann Ukhanova (DTU Fotonik, Denmark)
Kashaf Mazhar (KTH, Sweden)

Purpose of scalable video coding

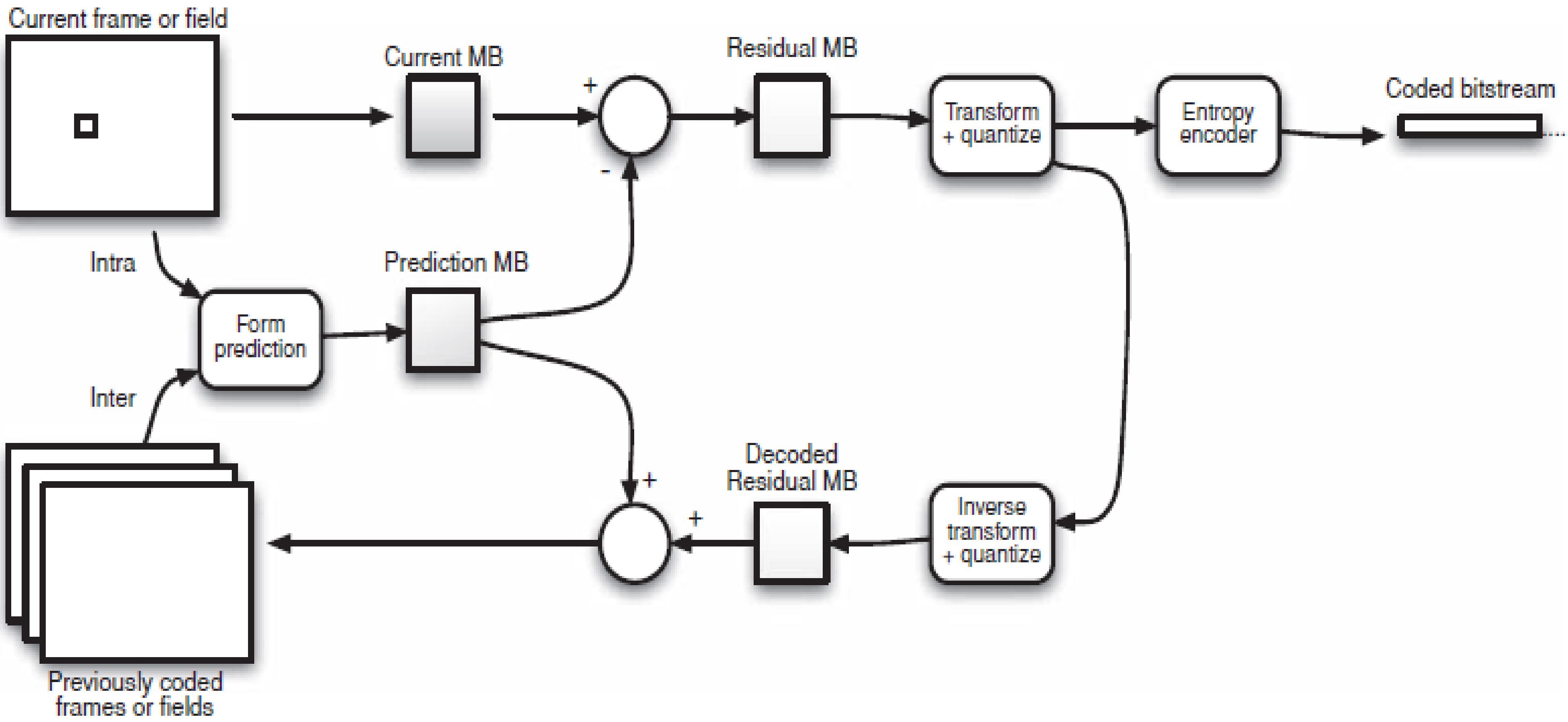
- Multiple video streams are needed for heterogeneous clients
- Scalable video stream - removal of parts of the video bit-stream to adapt to the various needs of end users and to varying terminal capabilities or network conditions
- Functionality of SVC
  - Graceful degradation when some parts of the bit-stream are lost
  - Bit-rate adaptation to match the channel throughput
  - Format adaptation for backwards compatible extension
  - Power adaptation for trade-off between runtime and quality

Introduction to H.264

- What is H.264 ?
  - H.264 is a Video Coding Standard also known as MPEG-4 Part-10 (AVC).

- H.261, H.263, MPEG-1, MPEG-2 are some predecessors of H.264.
- A Video standard specifically do not define an encoder; rather, they define the output that an encoder should produce
- A decoding method is defined

Structure of H.264 Codec

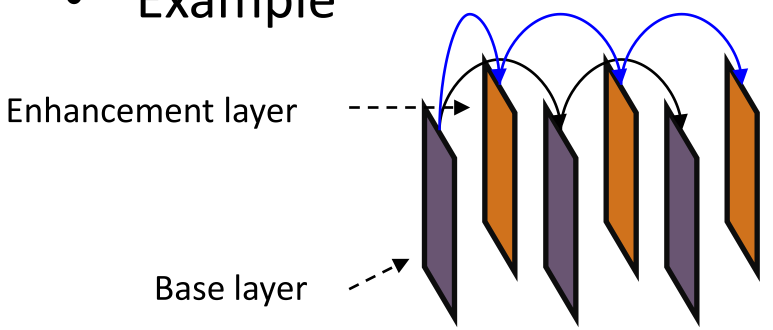


Scalability within H.264

- Video stream is encoded in more layers:
  - Base layer - minimum data required
  - Enhancement layers - refinements of the base layer
- Receiver can:
  - Decode all the layers for high quality video
  - Decode less layers according to its current bit rate
- Advantages :
  - Even with very limited bandwidth a user is able to receive a video stream
  - Uses less bandwidth by exploiting redundancy within a frame and between different frames

Introduction - modes

- Example



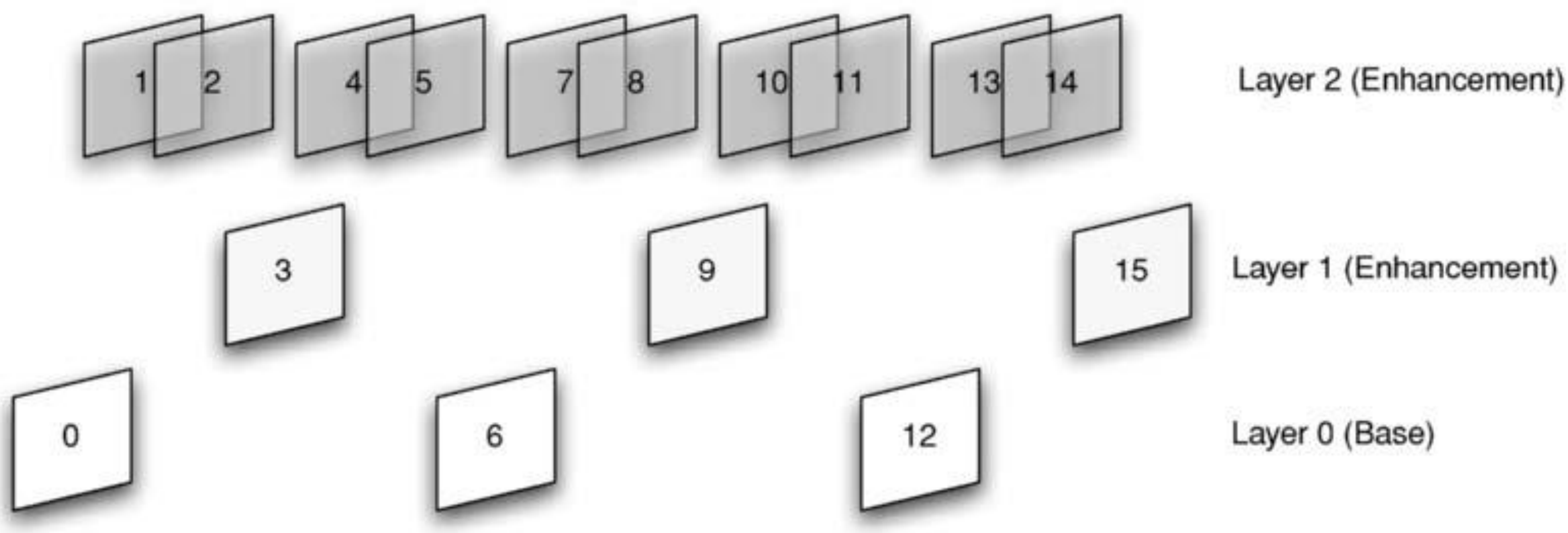
- Scalability mode
  - Fidelity reduction (SNR scalability)
  - Picture size reduction (spatial scalability)
  - Frame rate reduction (temporal scalability)
  - Sharpness reduction (frequency scalability)
  - Selection of content (ROI or object-based scalability)

Temporal Scalability

- Code a video sequence into several layers with the same resolution but with different frame rates

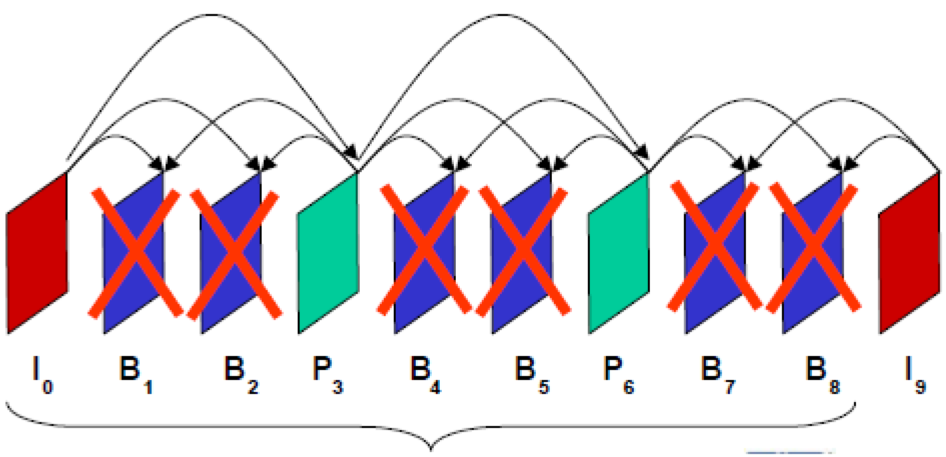
- The base layer has the lowest frame rate

- Enhancement layers add more frames to the decoded stream -> increasing the frame rate



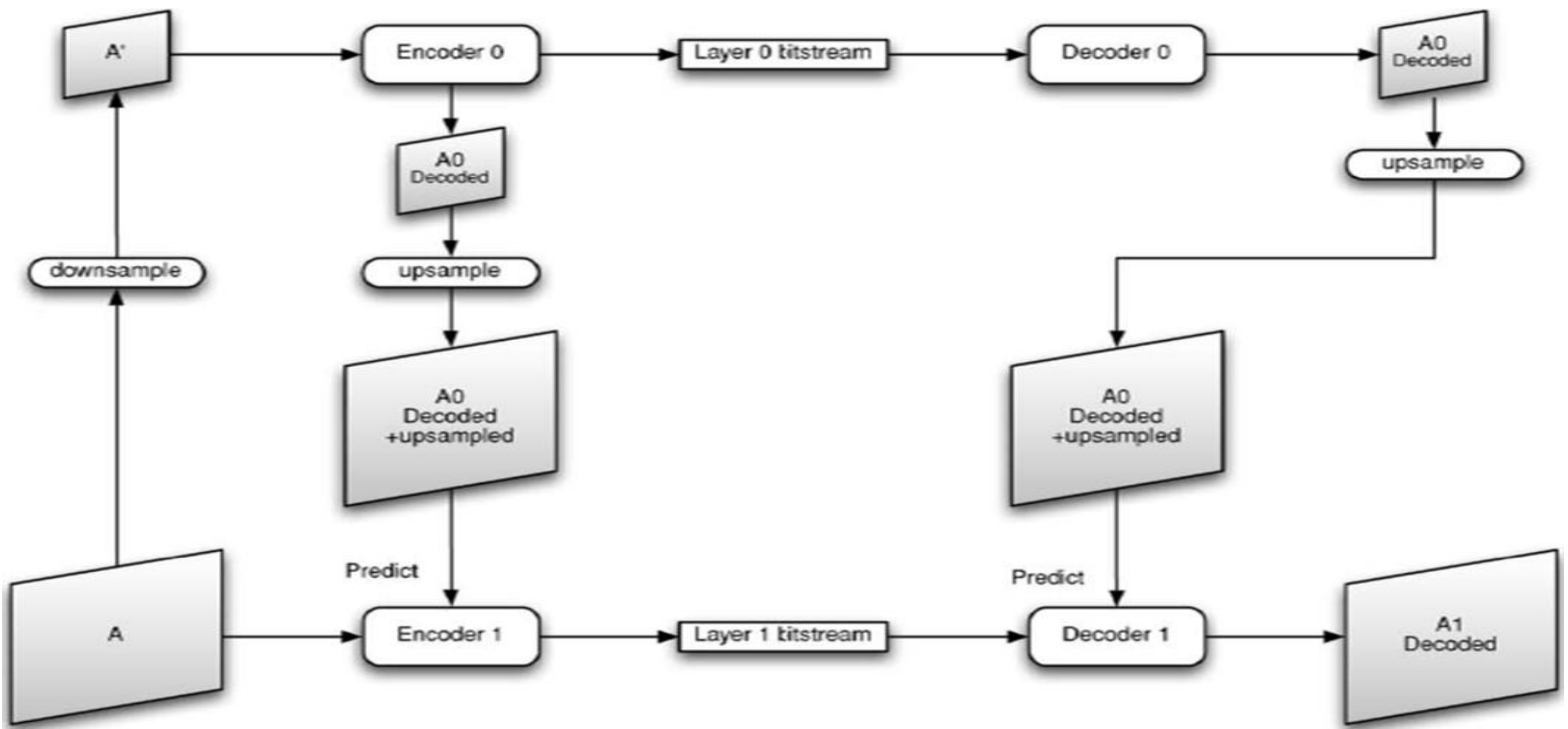
Temporal Scalability (cont)

- Based on the use of B-frames to refine the temporal resolution
  - B-frames are dependent on other frames
  - However, no other frame depends on a B-frame – each B-frame may be discarded without affecting other frames



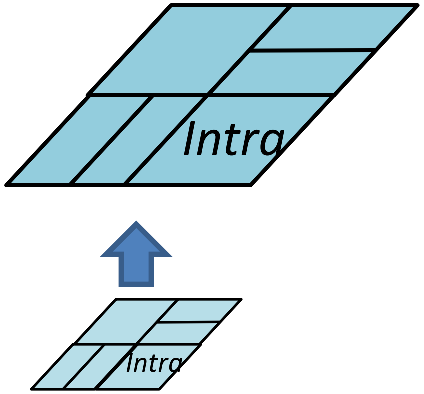
Spatial Scalability

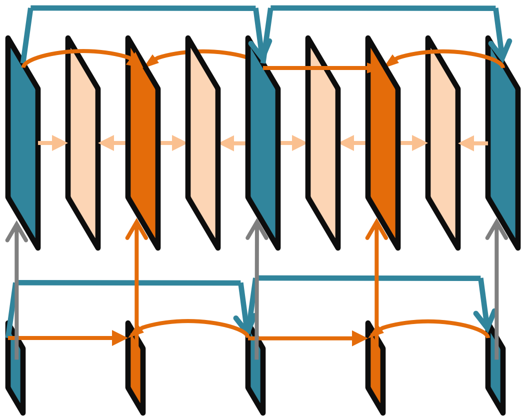
- Based on refining the spatial resolution
  - Base layer is low resolution version of video
  - Enhancement layer contains coded difference between upsampled base layer and original video
  - Also called: Pyramid coding



Spatial Scalability (cont)

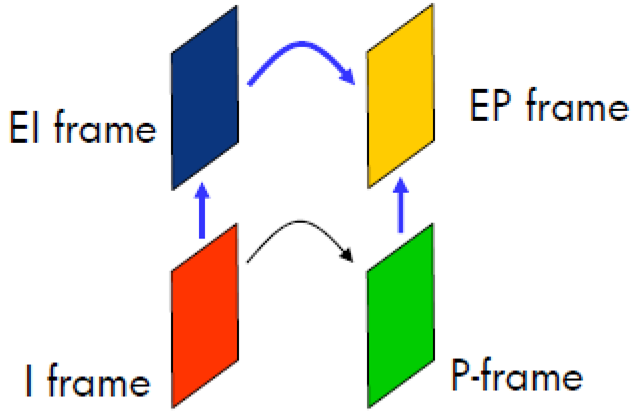
- Arbitrary resolution ratio
- The same coding order in all spatial layers
- Combination with temporal scalability
- Intra & Inter layer prediction





Quality Scalability

- Basic idea same as spatial scalability: Layer 0 has a higher QP than Layer 1
- Reconstructed frames from lower layers as prediction reference
- SNR (Quality) Scalability: Based on refining the amplitude resolution
- Base layer uses a coarse quantizer
- Enhancement applies a finer quantizer to the difference between the original DCT coefficients and the coarsely quantized base layer coefficients



Note: Base & enhancement layers are at the same spatial resolution

Objective Metrics

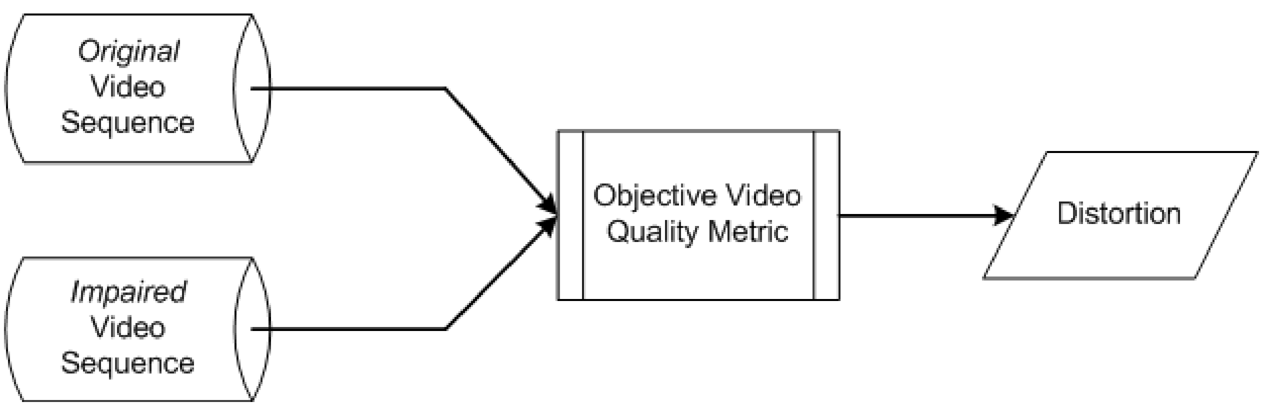
- Peak Signal-To-Noise Ratio (PSNR)
  - Used widely in evaluating coding performance
  - Purely mathematical difference

$$E_{RMS} = \sqrt{\frac{1}{M \times N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left[f'(m,n) - f(m,n) \right]^2}$$

$$PSNR = 20 \times \log_{10} \left(\frac{255}{RMSE} \right)$$

Video Quality Evaluation

- Objective
  - A computer algorithm judges the distortion between videos
  - Attempts to model a human observer
  - There is currently no standard method

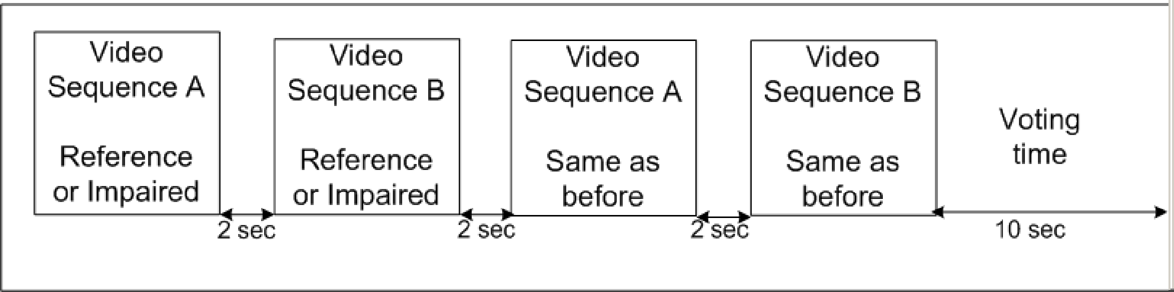


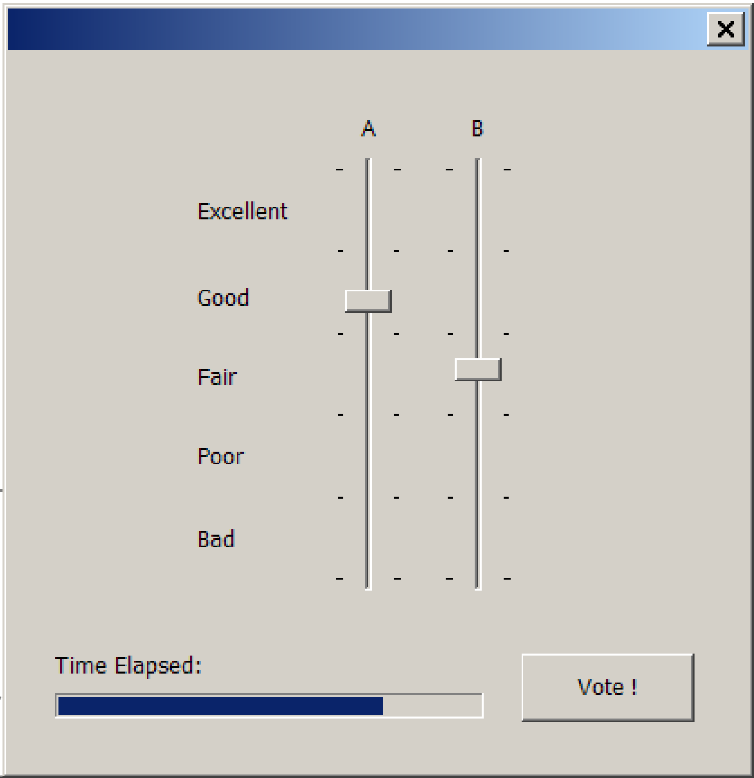
Video Quality Evaluation (cont)

- Subjective
  - A human "subject" rates the video on a scale

- Double Stimulus Continuous Quality Scale Method

  -- Difference is calculated as the actual rating





Experimental Study

- Project mainly encompasses the quality for different scalability options available for H.264/SVC

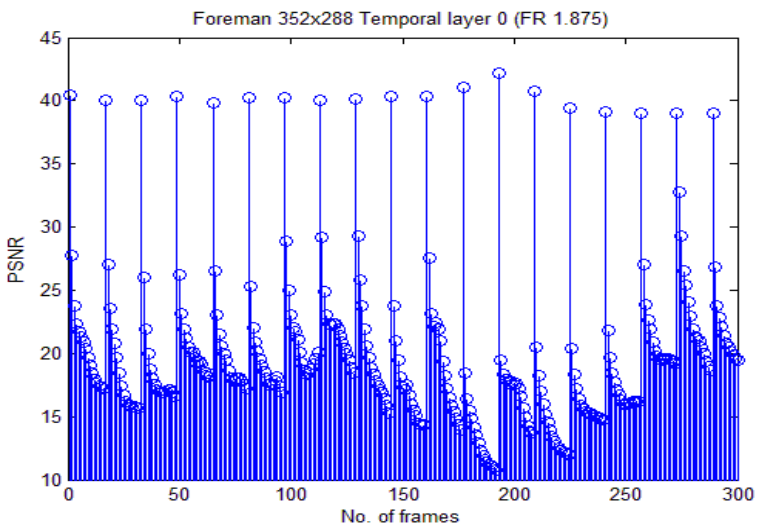
- Compress different video sequences (352*288, 30Hz)
- "akiyo"(news, static background)
- "hall monitor"(video surveillance, slow motion)
- "foreman"(interview, movie)
- "soccer" (football, fast motion)

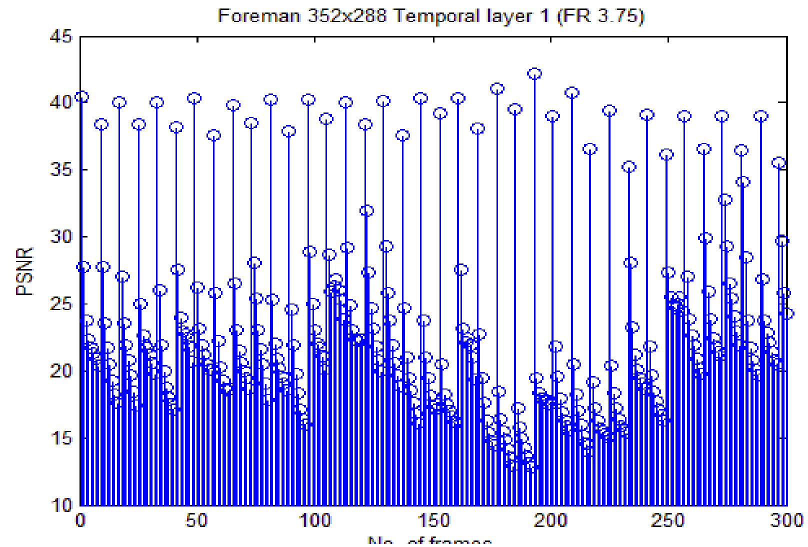
-- with all possible scalabilities – quality, spatial and temporal having different layers

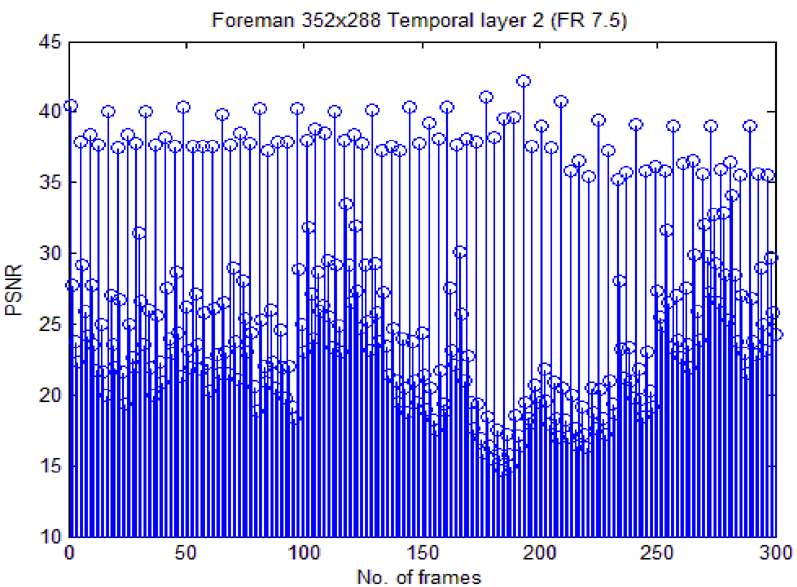
Challenge for Fair Comparison of PSNR

- Up sample and get the highest resolution(352x288) for spatially reduced video sequences
  - -- For all spatial layers with resolution of QCIF (176x144) each frame was up sampled by duplicating the Y pixels in the frame to match the resolution of the original video sequence
- Up-convert and get the highest frame rate (30Hz) for temporally reduced sequences (using frame interpolation and other methods)
  - -- For all the temporal layer with reduced frame rates the comparison of frame was made by duplicating frames to increase the number of frames and match the frame rate of the original video sequence.
- Keep the quality reduced sequences as it is, as we can not do anything to them

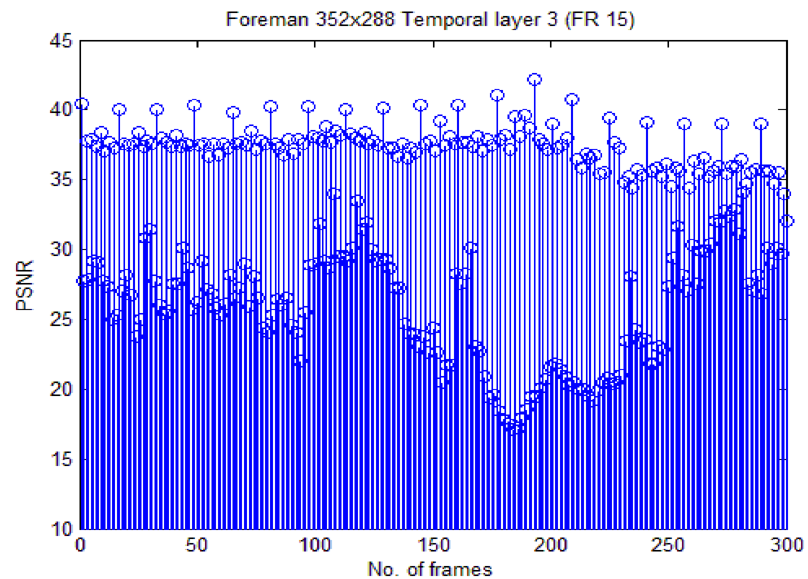
Temporal Layer PSNR Comparison for Foreman Sequence



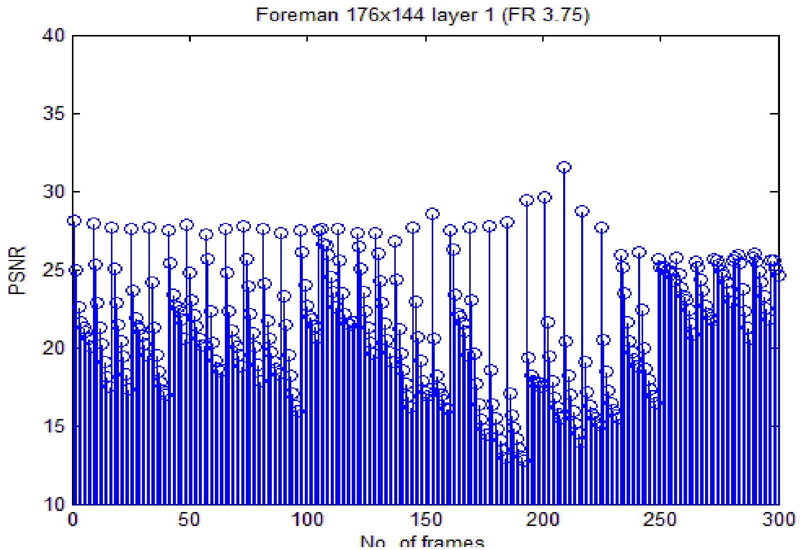


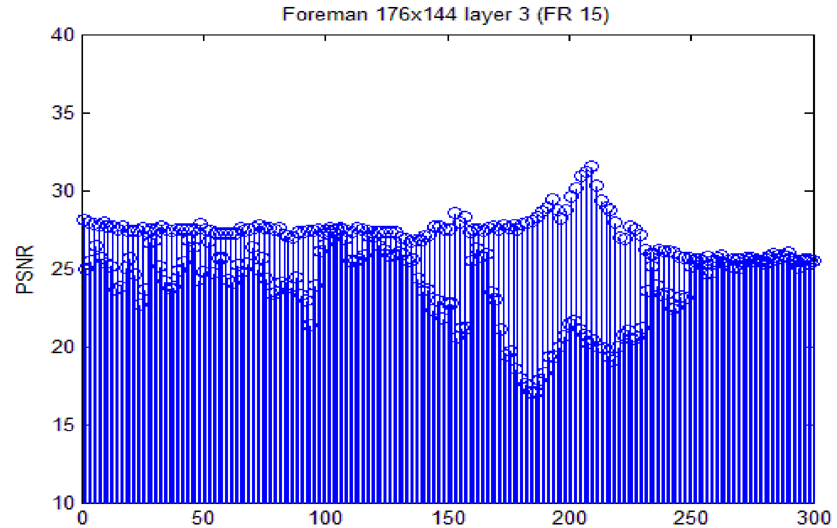


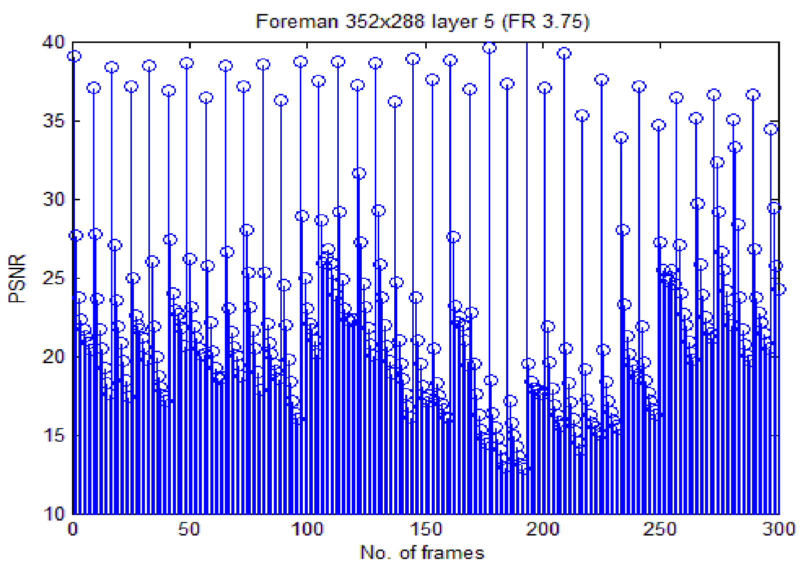
e, 1



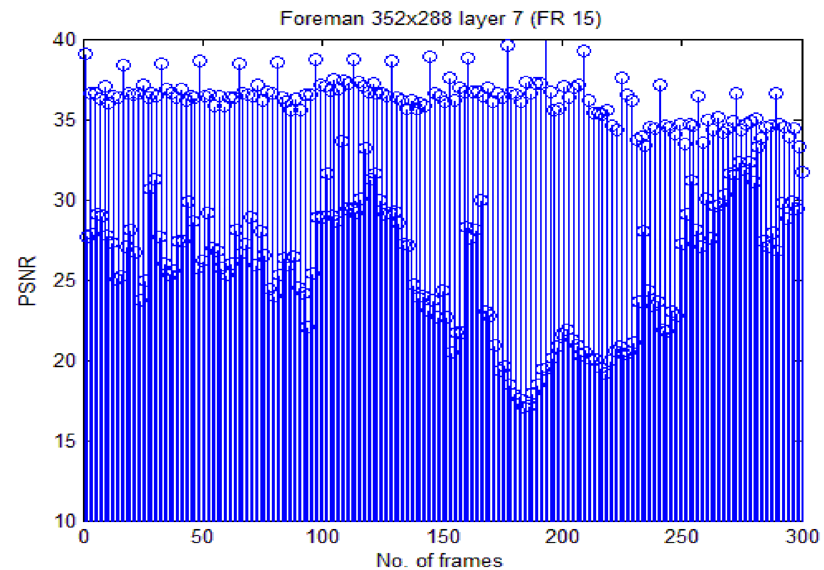
Spatial Layer PSNR Comparison for Foreman Sequence



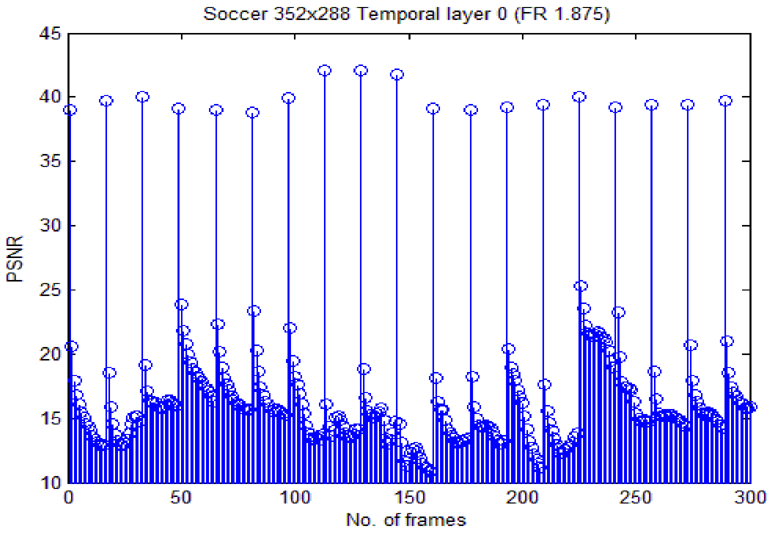


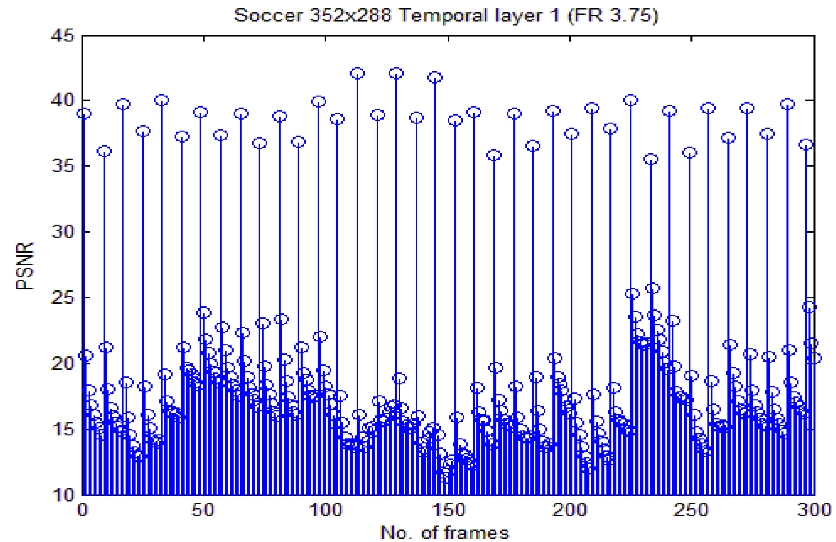


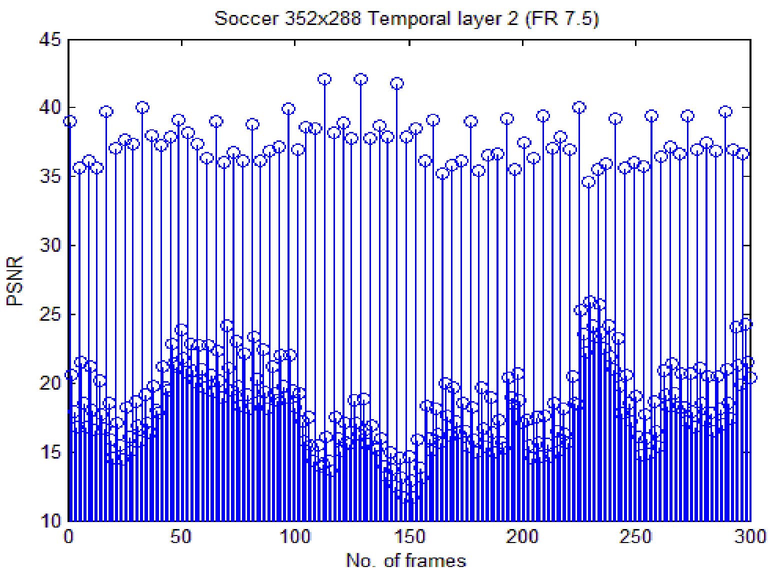
ence,



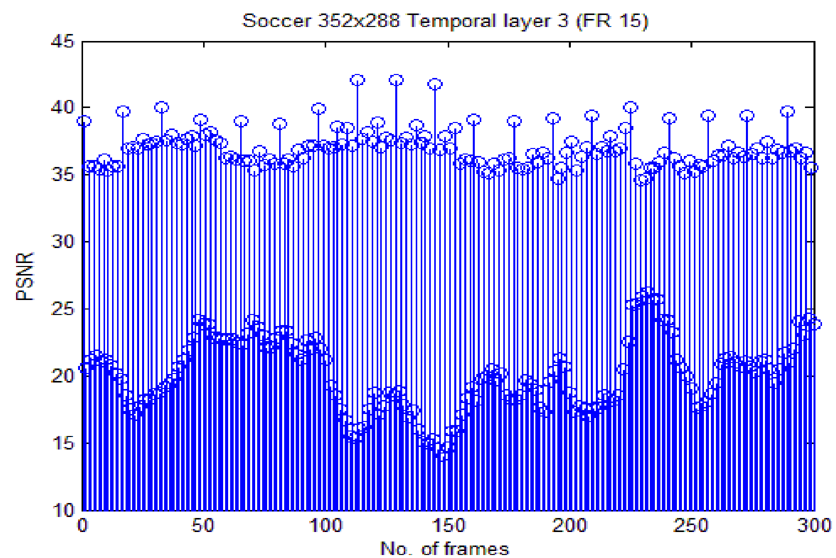
Temporal Layer PSNR Comparison for Soccer Sequence



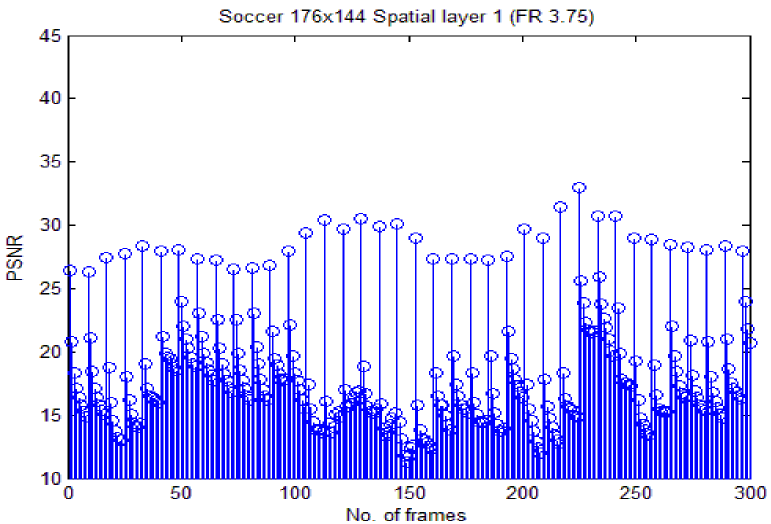


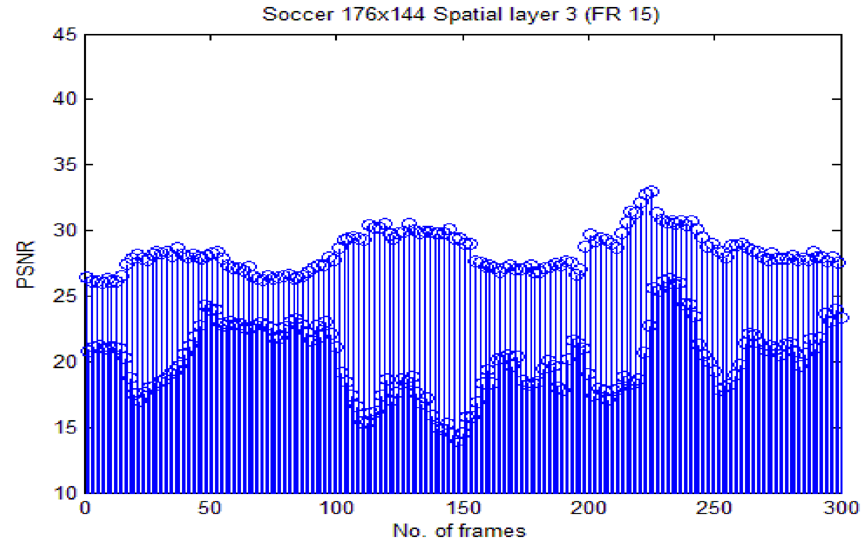


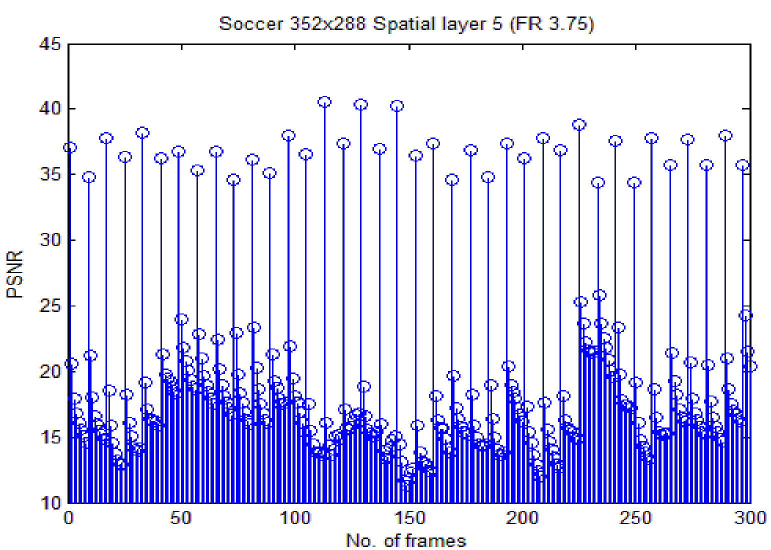
ference



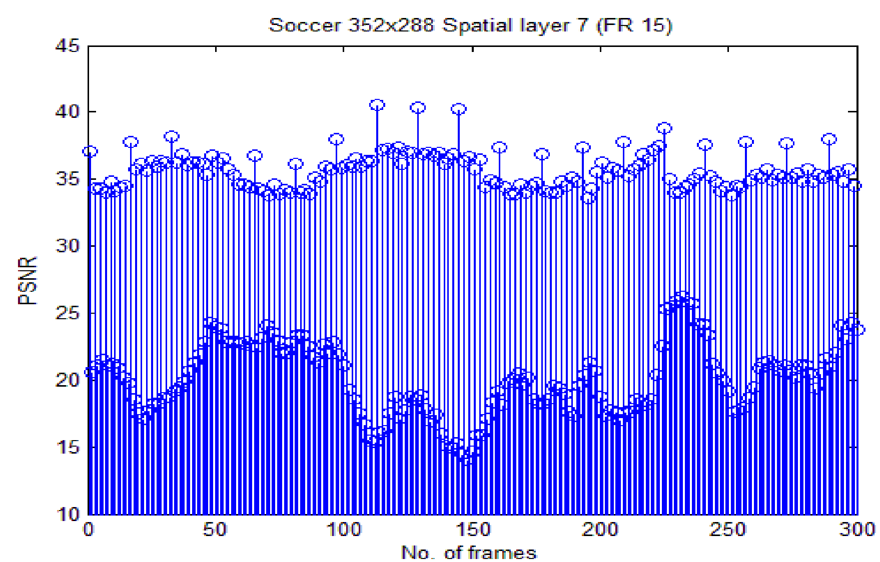
Spatial Layer PSNR Comparison for Soccer Sequence







fere



Analysis of Results

- As the FR improves for temporal scalability the PSNR improves.
- The low PSNR frames correspond to the duplicated frames for the fair comparison
- The comparison for PSNR for same FR spatial layers exhibit different values due to resolution of frames.
- Higher spatial layer with higher frame rates exhibit best PSNR
- For smaller FR the impact of resolution decrease is not so high as for higher FR. If the frame rate is low, for many frames for different spatial resolution the PSNR does not change a lot

Further work

- Subjective quality estimation
- Additional info about motion vectors etc.
- Model for scalable quality estimation
- Comparison with existing models

Questions?

Thank you!