CHAPTER 3

BAGBL FRAMEWORK

3.1 INTRODUCTION

Deaf and hard of hearing users are exceedingly benefitted from the computers and other technologies. The nature of internet is mainly visual and thus proves handy for the hearing impaired.

Sign language is a language which exploits unique features of the visual medium through spatial grammar. It relies on sign patterns viz., the movements and orientation of the arm and the body language of the person to facilitate understanding between people in which communication is viable without the means of acoustic sounds. It is evident that despite several handicaps, the number of people learning sign language has grown (Arun 2015). These include youth and corporate types keen on learning to effectively communicate with fellow deaf employees.

India, a country peopled with 17 million of such disabilities, it is estimated that only a few amongst them are familiar with the sign language (Arun 2015). This alarming situation is due to the less available teaching resources and the author hopes the dictionary developed by him for the same will help spread the word. Arun says that "International Sign Language (ISL) currently possesses around 5,000-7000 words, and as concepts improve in the deaf community they are added to the vocabulary," (Arun 2015). With less standard guidance and proper dictionary of the sign language, there is a
setback in the teaching of this language, any individual has to gather information from wide spread resources or from people with disabilities themselves. Due to these factors, this thesis works on the American Sign Language (ASL).

The rapid growth of smart phones and up-gradation of bandwidth in broadband networks has subsequently led to the expansion of mobile media traffic. A study by Cisco (2015) predicts that by 2015, two-thirds of world’s mobile traffic is to be predicatively multimedia. Quite the reverse, mobile media streaming is still an intimidating task, particularly for users in an extremely itinerant environment. The presence of heterogeneous access networks and high user mobility contribute to the wide fluctuations of wireless link qualities in terms of their throughputs and latencies. There is also a necessity that there should be an astute allocation of resources among competing traffic flows when the same access node (e.g., a cellular bases station or a WiFi access point) serves multiple video streaming sessions.

Sign language interpretation service of the video platform will face several confrontations by handling every virtually available Internet connection, across remote offices, small town hospitals and municipal bureaus. This also has to be extremely cheap and also easy to be used by people without any technical or computer skills. Encoding/decoding of real-time video and transmission of the same over network environments place a huge demand on processing of such data and their bandwidth capabilities. Most necessarily, adequate video quality is required to capture the dynamics of sign language’s fast-moving hand and finger movements. Wide link quality changes and dynamic traffic patterns force the individual users to have the most effective video streaming methods and this becomes the vital need of the day.
With the concept of layered video becoming more prominent & practical, efficient transmission of video over wireless networks with reduced variations are achieved using Hierarchical coding schemes like H.264/SVC (Scalable Video Coding) (Schwarz et al 2007).

This concept can be extended to gesture based streaming as well. Therefore, in order to facilitate quality transmission of gesture based video over wireless networks, this thesis proposes a BAGBL (Bandwidth Aware Gesture Based Layered) framework based on the concepts of Multiple Choice Multi-dimensional Knapsack (MMKP) problem (Lin 1998) and Pareto algebra (Geilen et al 2007).

Our established framework for gesture based video conferencing can be easily customized to integrate other multimodal based streaming applications as well. This framework will ensure that the video is adaptable to fluctuations of bandwidth in the network by dropping packets (layers). So, the expected quality can be downgraded under severe channel conditions. On contrary, as and when excess bandwidth is available, it tends to transmit all the layers efficiently and fairly by the optimized usage of the available bandwidth thereby improving the expected video quality.

3.2 BAGBL FRAMEWORK

This dissertation designs a BAGBL framework that consists of three major parts viz., a layered video representation, network with adaptive service, which uses two heuristics for bandwidth aware video streaming, and the receiver. Under the developed framework, as variations occur in the wireless channel, the video sender and the network elements can scale the video substreams and transport the scaled video streams to the receiver with acceptable perceptual quality.
3.2.1 Layered Video Representation

It is proven that for multimedia streaming services, a layered approach handles variation in bandwidth effectively and is a preferred method to achieve flexible bit stream adaptation in wireless networks.

A video sequence may be represented using an essential base layer (also called the main profile) and one or more optional enhancement layers (called scalability profiles). The base layer constructs the coarse or base representation of the stream, and the enhancement layers successively improve it. The base layer needs less transmission bandwidth due to its coarser quality; the enhancement layers require more transmission bandwidth.
due to its finer quality. The substreams are assigned priorities according to its significance. That is, the base layer is assigned the highest priority.

As stated, a shape energy based layered approach is proposed in the thesis that uses zernike moments to put the data of different importance into different layers. With such features this layered approach is more suitable for video transmission over an error prone channel with fluctuating bandwidth. Precisely, this concerns with generating a coded representation (bit stream) whose subsets appropriately reconstruct complete pictures of resolution or quality commensurate with the portion of the bit stream decoded. A base layer is obtained by decoding the minimum bit stream subset. Enhancement layers are generated by decoding the remaining bits in the bit stream to get more details of the video at higher resolution or quality as compared to base layer.

**H.264/SVC based layered representation**

Scalable Video Coding (SVC) is the successor of the H.264/AVC and is a very propitious encoding technique that adapts to streaming video over wireless networks with bandwidth fluctuations. SVC supports temporal, spatial and quality scalabilities at bit stream level that provides the means of easy acclimatization of video by selecting the subsets of bit streams. The pre-encoded SVC bit stream can very well be shaped by the streaming server for provision of a choice of spatial, temporal and quality (SNR) resolutions in proportion to the conditions of the wireless medium and capabilities of the receiver. In addition to this, the layered structure of SVC place the data into different layers according to their significance level. These features of SVC, prove that when video propagates over an error-prone channel with varying bandwidth, the scalable SVC bit stream is more suitable than the non-scalable bit stream (Schwarz et al 2007). We first use the standard hierarchical coding technique the H.264/SVC for generating the layered representation.
The raw YUV video is encoded with different encoding parameters (temporal encoding, spatial encoding, SNR encoding, or combined encoding) to generate an essential base layer (also called the main profile) and one or more optional enhancement layers (called scalability profiles).

Table 3.1 SVC Encoded Output

<table>
<thead>
<tr>
<th>Spatio-Temporal Resolution</th>
<th>Resolution</th>
<th>Bit Rate</th>
<th>Y-PSNR</th>
<th>U-PSNR</th>
<th>V-PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>176×144</td>
<td>49.45</td>
<td>34.58</td>
<td>39.78</td>
<td>41.49</td>
</tr>
<tr>
<td></td>
<td>176×144</td>
<td>68.43</td>
<td>33.81</td>
<td>39.52</td>
<td>41.16</td>
</tr>
<tr>
<td></td>
<td>176×144</td>
<td>92.03</td>
<td>33.33</td>
<td>39.40</td>
<td>41.01</td>
</tr>
<tr>
<td></td>
<td>352×288</td>
<td>174.72</td>
<td>34.52</td>
<td>40.89</td>
<td>42.66</td>
</tr>
<tr>
<td></td>
<td>352×288</td>
<td>231.26</td>
<td>33.29</td>
<td>40.57</td>
<td>42.32</td>
</tr>
<tr>
<td>EL1</td>
<td>352×288</td>
<td>301.93</td>
<td>32.51</td>
<td>40.41</td>
<td>42.15</td>
</tr>
<tr>
<td></td>
<td>352×288</td>
<td>392.89</td>
<td>32.72</td>
<td>40.46</td>
<td>42.20</td>
</tr>
<tr>
<td></td>
<td>704×576</td>
<td>603.85</td>
<td>35.56</td>
<td>42.46</td>
<td>44.64</td>
</tr>
<tr>
<td></td>
<td>704×576</td>
<td>806.20</td>
<td>34.78</td>
<td>42.46</td>
<td>44.31</td>
</tr>
<tr>
<td></td>
<td>704×576</td>
<td>1068.02</td>
<td>34.33</td>
<td>42.07</td>
<td>44.18</td>
</tr>
<tr>
<td>EL2</td>
<td>704×576</td>
<td>1384.10</td>
<td>34.03</td>
<td>42.00</td>
<td>44.13</td>
</tr>
<tr>
<td></td>
<td>704×576</td>
<td>1637.61</td>
<td>33.97</td>
<td>42.02</td>
<td>44.13</td>
</tr>
</tbody>
</table>

Table 3.1 shows the scalability of H.264/SVC with one base layer and two enhancement layers with substreams inside each. Here, the base layer (BL) gives Quarter Common Intermediate Format (QCIF) video with a resolution of 176×144 and specified frame rates. The enhancement layer (EL1) improves video to Common Intermediate Format (CIF) with a resolution of 352×288. The Enhancement layer (EL2) is of an enhanced quality with a resolution of 704×576. This feature of SVC encoded videos is applied for streaming by just dropping the least important levels depending on
the available bandwidth and client requirements. SVC is not as efficient as AVC encoding. Moreover, because of a lack of hardware decoding support, SVC has not been used as much. Therefore, a layered approach is proposed which uses shape energy trajectory of hand sign gestures for video layering.

**Shape energy based layering**

Applications such as the video analytics, which suffer from the issue of dealing with processing a huge number of video frames use key frame extraction as a pre-processing step (Momin & Pawar 2014). This concept can be applied for video streaming applications as well. "Key frame extraction, is the problem of finding the minimal set of key frames that cover all significant events or maximize the number of key frames while minimizing redundancy of information in these key frames"(Ejaz et al 2012). Video layers are generated by arranging the video frames according to the amount of information present. Video streaming is achieved by just dropping the least important levels (layers) depending on the available bandwidth and client requirements.

**Key Frame Extraction using shape energy trajectory of gestures**

"Video summary is a concise and informative representation of the source video". The two most prevailing approaches of generating video summaries are static and dynamic, even though several other methods are available for the same.

The static video summary concerns about extracting the so-called key frames from the video. The key frames are nothing but the still frames extracted from any sequence which holds the most significant or substantial content of that particular video sequence and thus serves as a representative of that video sequence.
The dynamic video summary concerns about accumulating small shots in a time ordered sequence. Dynamic video summary is advantageous over the static summary in the sense that it retains the dynamic nature of video content through positioning of the semantics pertaining to time. In addition, the dynamic summaries propagate impressive information due to the audio and motion contents.

Key-frames are extracted based on the curvature of the “shape energy” trajectory to create a content–based representation for them to select only those frames that are able to describe adequately the performed gestures.

![Figure 3.2 Key Frame Extraction](image-url)

Figure 3.2 Key Frame Extraction
Pre Processing

Pre-processing is an important step for extracting possible choices or candidate frames from the input video sequence. This includes various steps starting from segmentation, masking to hand detection.

To start with the hand is to be segmented. There are various approaches of hand segmentation, among which we choose to work with the skin colour for detecting and segmenting hands (Jones & Rehg 2002) even though it is not a reliable modality by itself. It is assumed to have a uniform background and clothes for simplification.

A threshold is used to perform segmentation. After that, the segmented hand (binary image) could be used as a mask, for extracting the Zernike moments as shown in Figure 3.2. Zernike moments are orthogonal and so have better representation capabilities than other moments like Cartesian or Hu moments. Therefore, the thesis chooses to work with Zernike moments. Due to the principle of orthogonality, the sum of the squared coefficients of the Zernike moments expresses the “energy” of the gesture shape which is used for layering purpose in our framework. The above process is elaborated as follows:

Skin detection using color information involves representation of the image pixels in a suitable color space (the choice of the color space involves several factors), after which the skin and non-skin pixels are modeled using a suitable distribution and finally classifying the modeled distributions.
Color Spaces

The thesis works with the HSI or HSV color space. RGB (Red Blue Green) cannot represent the perceptual features of color such as Hue (H), Saturation (S) and Intensity (I) in a straight away manner, and therefore to map RGB on to the perceptual features several non-linear transformations are proposed. The HSV (Hue Saturation and Value) space defines color as Hue—the property of a color that varies in passing from red to green, Saturation—the property of a color that varies in passing from red to pink, brightness (also called intensity or lightness or value)—the property that varies in passing from black to white. It is proved that the transformation of RGB to HSV is invariant to high intensity at white lights, ambient light and surface orientations relative to the light source and hence, can form a very good choice for skin detection methods (Kakumanu et al 2007).

Skin Color Classification

In the context of classification, skin detection can very well be debated as a two class problem viz., skin-pixel vs. non-skin pixel classification. Among the variety of methods used for skin classification, Gaussian mixtures can be used for skin-color modelling due to the most important feature that these parametric models can generalize well with less training data. It should also be noted that they require a very less repository space. The skin colors of different individuals accumulate in a very tiny space under controlled illuminations. Therefore, it is concluded that, under certain lighting conditions, the skin-color distribution of different individuals can be modelled by a multivariate normal (Gaussian) distribution in normalized color space.

To recapitulate, a face detector from an obtainable vision library helps us in locating the face region and then based on the face color we train a
color classifier. Proceeding further, the hand regions are detected from the modelled color as detailed further:

We assume a single multivariate Gaussian model whose probability density function in the HSI color space is given by

\[
p(x \mid \text{skin}) = \frac{1}{\sqrt{(2\pi)^3 | \Sigma |}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}
\]

where \( x \) refers to the color vector and \( \mu, \Sigma \) are the mean vector \( \mu \in \mathbb{R}^n \) and the covariance matrix \( \Sigma \in S_{++}^n \) respectively of the vector valued random variable \( X = [X_1 \ldots X_n]^T \). By calibrating the skin probability for all pixels we acquire the skin regions. Now the hand image segmentation of skin and non-skin regions using the color model results in a binary mask.

As grey-level images provide a better representation of the current gesture than the binary masked one (as it may lead to loss of information especially in the case of different gestures with similar silhouettes, e.g., front and back image of the hand), we obtain the grey-level image of the masked one by applying it to the Intensity channel. The above is performed for only the hand regions and the head.

As stated earlier, due to the reduced information redundancy and reconstruction capability, Zernike moments are used to represent the activity state which is a representation of the relative position and shape of the head and the two hands.

**Key frame extraction**

Now, after performing all sorts of pre-processing, we extract the key frames based on the curvature of the “shape energy” trajectory to create a
Let $V_t$ denote the original video sequence at an instant “t” captured at a rate of “$N_f$” frames per second represented by a set of frames “$F(t)$” described as $V_t = \{F(t)\}$. The aim is to derive a captured set of key frames $KF_t = \{F_k(t)\}$ from $V_t$ such that the gesture energy “$J$” represented by the sum of the squared coefficients of the Zernike moments is maximum.

$$J = \sum_{n=0}^{n_{\text{max}}} \sum_m ||Anm||^2$$

Where $n_{\text{max}}$ is the selected order of the moments. Here $n_{\text{max}}$ is taken as 2, the second derivative of $J$ denoted by $J''$ which gives the curvature measure. Local maxima correspond to time instances of peak shape variation while local minima indicate low shape variation.

Let “$X_M$” contain the time instances of frames corresponding to the local maxima of $J''$ estimated as $X_M = \{F(t) : J''(F(t)-1) < J''(F(t)) \& J''(F(t)) > J''(F(t)+1)\}$ and “$X_m$” contain the time instances of local minima of $J'$ estimated as $X_m = \{F(t) : J'(F(t)-1) > J'(F(t)) \& J'(F(t)) > (F(t)+1)\}$.

$X_M$ that contains the local maxima is denoted as the set of key frames extracted from the video sequence, $KF_t = \{X_M\}$.

**Layering**

The base layer is formed by the second order moments of the extracted key-frames which preserve the semantics of the gesture. The higher order moments contribute to the enhancement layers. At the client side, the image can be reconstructed using a set of moments through order “$M$” as given by
\[ \hat{f}(x,y) = \sum_{n=0}^{n_{\text{max}}} \sum_{m} A_{nm} V_{nm}(\rho, \theta) \]

where \( \hat{f}(x,y) \) is the reconstruction of the original image \( f(x,y) \), “\( n \)” is a positive integer, “\( m \)” is an integer such that \( |m| \leq n \), and \( (n - |m|)/2 = 0 \), and “\( n_{\text{max}} \)” is the maximum value of \( n \) (maximum order). As higher values of \( n_{\text{max}} \) capture a lot of noise, \( n_{\text{max}} \) is determined by computing moments up to an order where there is only a 1% difference between the reconstruction and the original.

### 3.2.2 Bandwidth Aware Video Streaming

If the bandwidth available during transmission is less than the required bandwidth for transmission of the packet, the packet may obviously be lost. Moreover, if the sender is unaware of the channel status, and tries to transmit packets, then there is a possibility that all layers be discarded with equal probability, which results to poor video quality. In order to overcome this, by taking into account the network conditions, the dissertation proposes a method to pre-emptively discard enhancement layers in an intelligent manner at the sender side.

Streaming refers to the transmission of different video layers in different bit streams, called substreams. As discussed above, multiple layers or substreams are generated which is then fed to the network. Depending on the availability of resources in the wireless medium, the bandwidth aware streaming subsystem is to scale the substreams for transmission through the medium.

Specifically, the bandwidth aware streaming system includes the following functions:
A reservation of minimum bandwidth to transmit the base layer leads to an acceptable reconstruction quality.

Transmit or discard, i.e., adjust the enhancement layers based on the bandwidth availability considering fairness. That is, the video streams are scaled depending on the availability of the resources.

The streaming subsystem is divided into two major parts, viz., Substream Scaling, or Substream Selection, and Substream Scheduling.

**Substream Selection**

A scaling function is incorporated in the framework to act during bandwidth fluctuations as indicated: During mobility, as the available bandwidth on the path reduces, the scaler(s) drop the lower priority substreams and transmit the higher priority substreams, and as bandwidth on the path increases, resulting in the perceptual quality improvement/increase at the receiver.

A heuristic is devised (MPMHSS) based on the concepts of Pareto minimisation and MMKP for solving the problem of substream selection.

The base station manages the process of notifying the sender through a signalling channel, the current network/medium status by collecting the information, like the current available bandwidth with the help of a bandwidth manger. Therefore, the bandwidth manger is responsible for the decision on scaling.

Thus, the scaler in the selection module filters the layered video by selecting certain video layers for transmission by dropping certain layers with respect to their significance. The order of dropping is from the highest enhancement layer which has the least preference/priority down to the base
layer that obviously has the highest priority or preference. The bandwidth manager decides on the scaling process according to the information received by the bandwidth estimator. A heuristic (MPMHSS) is devised based on the concepts of Pareto minimisation and MMKP for solving this problem of substream selection, which is discussed in detail in Chapter 4.

**Substream Scheduling**

Once the substreams are selected, the packets are to be scheduled for transmission over the medium with respect to their priorities and QoS specifications. Depending on the wireless channel conditions, there is a possibility that a lower priority substream is dropped for a period of time, for transmission of enhanced priority substreams. Thus, a scheduler provides a scaling function as well; however, its scaling function is a result of its scheduling function.

The scheduler reacts to the fluctuations in the channel with the feedback received taking into account, a substream’s QoS specification, the relative importance between the substreams and the wireless channel conditions.

To achieve this, the Dynamic Packet Scheduling algorithm (Nasser et al 2013) is slightly modified and a heuristic called Modified Dynamic Packet scheduling algorithm (MDMP) is devised, where a queue is formed with the base layer containing the highest priority and the enhancement layers assigned subsequent priorities. Also, it is assured that, at any instance, all the queues should contain a base layer and its corresponding enhancement layers. To enforce this, a sliding window is assumed which is formed by one base layer ($L_0$) and its corresponding enhancement layers ($L_i$). All the packets are sent in a non-pre-emptive fashion and it is assumed that the current available bandwidth of the network is always greater than the data rate of the base layer.
packets guaranteeing their transmission in order to ensure a minimum
received video quality and depending on the bandwidth availability,
additional enhancement layer packets are transmitted.

3.2.3 Receiver

The reconstructed video at the receiving end is passed through the
sign language recognition system, which uses filtering, skin color detection,
and hand-postures-contours comparison algorithms for detecting face and
hand regions, and a systematic process of principal component analysis along
with Sober edge detection for identifying the hand gestures. To recognize the
sign gestures, Hough transform and Feed Forward back propagation network
is used. Finally a Verbose Text to speech software is used to convert the
resultant text to speech. Analysis showing the recognition rate and the PSNR
of the received video is dealt with in detail in Chapter 4.

3.3 SUMMARY

A bandwidth aware gesture based video streaming
framework/architecture is proposed, which is targeted to support transport of
quality gesture based video over mobile wireless networks. Under the
proposed framework, as and when the characteristics of the wireless medium
changes, the sender and the network will be able to adapt to the same by
scaling the video streams thereby achieving adequate perceptual video quality
at the receivers. Thus, our proposed framework achieves efficiency and
fairness with good recognition rate for the gesture video.