Chapter 9
Conclusions and Future Directions of Research

The analyses of the methodologies adopted for the repackaging and the subsequent content re-creation of instructional video for technology enhancement in distance education have been provided in previous chapters. The conclusions derived out of this research are summarized in this chapter along with highlighting the novelties, merits and shortcomings in this research area. Some future directions of this research are also included towards the end.

9.1 Conclusions

The preliminary tasks addressed in connection with the repackaging of lecture video are shot detection and recognition. With this objective, two different approaches have been suggested. In the first one, the histogram difference in the intensity space temporally segments the video and an HMM classifier identifies the scenes subsequently. In another approach, the HMM framework designed for the shot classification itself performs the change detection, through a continuous monitoring of the respective likelihood functions using Shiryaev-Robert statistics. This framework is capable of automatically detecting the transitions in scenes, thereby separating them at the transition points so that the individual activities can be efficiently and continuously recognized with a guaranteed minimum delay. Using these methods the instructional activities are classified into three categories: (1) talking head, (2) writing hand and (3) slide show. We give a nomenclature of content frames to both the second and the third classes as a whole and non-content frames to all frames belonging to the class of talking head.

The first set of investigations in this research is focused on the development of a simple but accurate algorithm for compression efficient, content preserved repackaging of lecture video sequences. In this, computation of visual quality and content based key-frame extraction strategies are applied differently on the separate classes of video segments and a pedagogic representation of these lecture video segments is done effectively. For the selection of key-frames in non-content segments like a
talking head, a no-reference perceptual visual quality measure based on the blocking effect and blur and as recommended in [154] is used while for content segments, an HPP based scheme is employed. The Radon transform based skew detection module avoids duplications in content key-frames and yields distinct key-frames. The key-frames for different scenes of the lecture video provide a visual summary and an effective description of the content of the lecture. If the content key-frames suffer from poor resolution, they are super-resolved for improved legibility. A MAP based technique is used to combine frames temporally adjacent to a key-frame and super-resolution of the content frame is achieved. On media synthesis, an estimate of the original lecture video sequence is re-created from these content key-frames, the talking head frames and the associated audio. The reproduced media completely represents the original video in terms of all pedagogic values since it is delivered to the viewer in a semantically organized way, without altering the playback duration of the video. The multimedia summary produced by the content analysis which essentially contains only a few key-frames, a text file with the temporal information as the meta-data and the associated audio can be called an *instructional media package* (IMP). The summarization and the content re-creation of lecture videos as described in Chapter 5 lead to considerable savings in memory when an enormous amount of video data are to be stored. This, in turn, accounts for a fast browsing environment, on a huge volume of lecture video data. Another achievement that can be noted is the reduction in the requirement of transmission bandwidth.

The next part of research is an extension of the above work which aims at developing a method for legibility retentive display of instructional media on miniature display devices that tries to preserve the pedagogic content in the original instructional video. Thousands of video frames have been represented by a few corresponding key-frames which are displayed on the miniature screen as moving key-hole images. The size of the key-hole image is comparable to that of the mobile screen and so only the *region of interest* (ROI) in the key-frame with respect to the audio is displayed with the full resolution. The movement of the key-hole image is controlled by a video meta-data derived from the tracking of the written text in the content segments of the video. This meta-data is derived by using both the HPP and VPP of the ink pixels in the content frames. Originally the meta-data is computed in 50 frame intervals and on display on mobile devices, these are temporally interpolated to yield a faster frame rate for smooth panning. This method does not crop windows from the resident key-frames in the memory. Instead, it selects the rectangular region of interest in accordance with the tracking meta-data and delivers this region on the mobile screen on playback. In the default case, this panning window moves automatically with the tracking meta-data. One may also have a manual control through the touch screen or the selection key on the mobile keypad. Thus the movement of the window can be additionally controlled manually as per wish of the viewer if, for some reason, the hand tracking meta-data performs poorly. Moreover, the visual content delivery is performed not always with the single (fullest) resolution. If the text tracking meta-data variation over a short interval in time is much more than the size of the display window, the algorithm expands the region of interest but then the visual delivery will be at a reduced resolution. This helps in