# Automatic Capsule Preparation for Lecture Video

A. Ranjith Ram and Subhasis Chaudhuri
*Vision and Image Processing Laboratory*
*Electrical Engineering Department*
*IIT Bombay, Powai*
{*ranjiram, sc*}*@ee.iitb.ac.in*

***Abstract*—In this paper, we propose a visual quality and content based approach for creating a lecture video capsule which essentially contains the highlights of the original lecture video. We first segment and recognize the activities in the instructional video using a hidden Markov model(HMM). The activities are classified into three categories : talking head, writing hand and slide show. Then a no-reference objective quality assessment of the non-content frames, i.e., talking head frames, are performed to detect a few high quality frames for highlight creation. A new method of defining the visual quality of content frames, i.e., writing hand and slide show frames which are very crucial in a lecture video, is proposed in this paper. Statistical parameters of the intensity histogram along with the horizontal projection profile(HPP) of the the content frames are used to derive an objective quality measure of these frames. This allows us to extract the high quality content frames, which are 'well-written'. Finally on media re-creation to produce the video capsule, we select the video segments of a few seconds duration around these good quality frames along with the audio and merge with suitable proportion of the detected classes of instructional activity. This technique accounts for a very compact representation of the whole lecture video which finds a very potential application of video previews.**

***Keywords*-HMM; HPP; content; highlight; video-capsule;**

## I. Introduction

Making a long video short is still a promising field to video researchers. It is due to the fact that recently the world wide web experienced an increased use of digital video which opened new dimensions in the field of education and entertainment. Consequently, research and development of new multimedia technologies which will improve the accessibility of the enormous volume of stored video are inevitable. In the field of education, especially where the potential users are students, a fast preview of the entire lecture would be very useful for them before attending or even buying the video for the whole course [1, 2]. In this context, a lecture video capsule will help them for a fast preview of its content. This is the motivation of our work.

The lecture video capsule should contain the highlights of the entire lecture. Although there are many works reported on the highlight creation of sports video [3, 4, 5], that of an instructional video is still an ill-addressed one. In the case of sports video, interesting events occur only with increased crowd activity by which the audio energy level is boosted much. Hence audio is an important cue for sports video highlight creation. Also, visual information of the desired highlight, for eg., a goal hit or a boundary hit, vary much from the ambient one, in sports video. Hence almost all of the methods for sports video highlight creation use predefined models which make use of visual as well as audio signal. In [3], authors use audio, text and visual features for the automatic extraction of highlights in sports video. In [4], authors detect highlights using audio features only without relying on expensive visual computations. In [5], authors build statistical models for each type of scene shots with product of histograms and then an HMM is learned for each type of highlight. In [6], authors present an HMM based learning mechanism to track video browsing behaviour of users which is used to generate fast video previews.

Since an instructional activity is shot inside a classroom, its features are much different from sports or commercial video. The audio energy level is more or less same throughout the lecture and the only cue that can be used, is the visual information. Moreover, domain models cannot be employed for this because the highlight occurs at some point of the same activity like a talking head, writing hand or slide show. Hence a new strategy is to be employed for defining the highlights in lecture video. We base our approach on two observations. The first is that in the case of non-content segments like talking head, mere visual quality [7, 8] will suffice to create highlight. The second is, in the case of content segments like writing hand or slide show, a quality measure based on both the visual quality and content is to be used to extract the instructional highlights. Hence for defining the quality of the content frame, we adopt a method which utilizes statistical features of the visual content along with the HPP of the frame. The high quality clips from both the content and non-content segments are used in suitable proportions, to create the video capsule. Audio coherence is maintained on selecting these clips during media re-creation.

The rest of the paper is organized as follows. In Section II, we present the proposed method of video capsule preparation with suitable subsections. In Section III, we discuss the results of shot detection and recognition, visual quality assessment and media re-creation. Section IV summarizes our work with some concluding remarks.

## II. Proposed method

A block schematic of the proposed method is given in Fig. 1. It works with the following distinct steps for the automatic capsule preparation of lecture video:

(a) Shot detection and recognition

International Workshop on Technology for Education (T4E), Aug 4-6, 2009, Bangalore

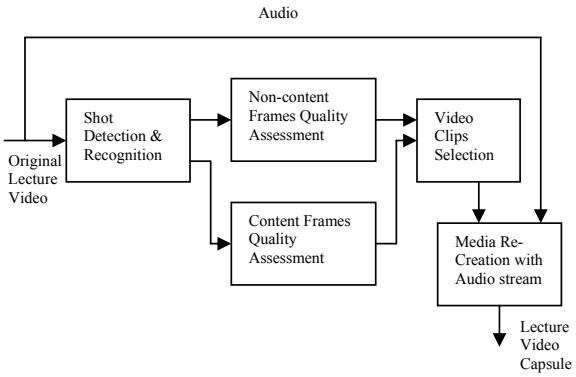



Figure 1. Block schematic for the proposed method of capsule preparation of lecture video.

(b) Visual quality assessment of non-content frames
(c) Visual quality assessment of content frames
(d) Media re-creation for capsule preparation

Since a lecture video generally consists of sequences of talking head, writing hand, slide show, audience (discussion), demonstration (imported video) etc., the first and fundamental step in the analysis should be the temporal segmentation to detect the scene changes. Then the activities in these segments are to be recognized. The writing hand and slide show frames are called content frames and others, non-content frames, which are depicted in Fig. 2.

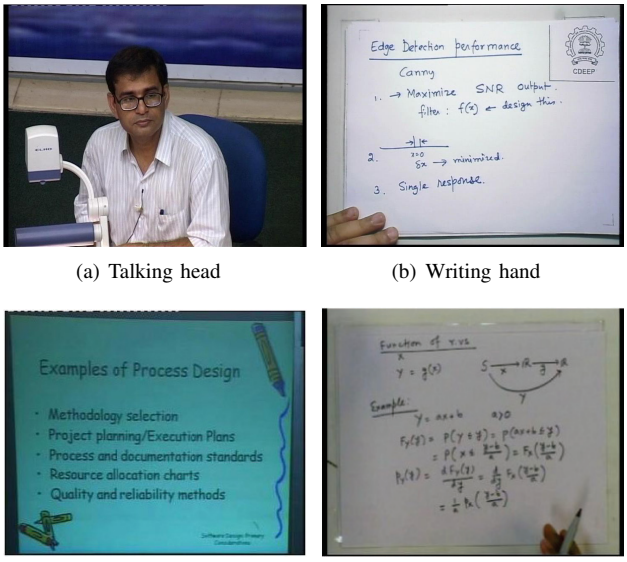
### A. Shot detection and recognition

The histogram difference is used for the temporal segmentation. It is given by

$$D(t) = \sum_{i=0}^{255} \mid h_t(i) - h_{t-1}(i) \mid \qquad (1)$$

where $h_t$ is the normalized intensity histogram of the current frame and $h_{t-1}$ is that of the previous frame. To improve the speed of computation, the original 256-level histogram is converted to a 16-bin histogram and then processed. If the sum of the absolute difference of the histograms, expressed in equation (1) crosses a threshold, corresponding frame is declared as a shot-boundary frame. This histogram difference method works well, as there is very little camera movement and hence, a simple but fast scene change detection can be performed.

Since HMM [9, 10] is a powerful tool for speech or activity recognition, it is used in our approach for shot classification. Several works were reported on HMM based activity detection in video [11, 12, 13]. The HMM based activity recognition has two phases, namely training phase and testing phase. The frame sequence $I = \{I_1, I_2, ..., I_T\}$ is transformed into observation sequences $X = \{X_1, X_2, ..., X_T\}$ for the learning and recognition phases of HMM, where $X_n = \{x_{1,n}, x_{2,n}..., x_{D,n}\}$, in which $x_1, x_2, ...$ are the features and $D$ is the dimensionality ($D = 2$ in our algorithm). In the training phase, for each class of instructional activities, a feature vector $x_i \in \mathbb{R}^n$ is extracted for each frame $I_i$ and *pdfs* are constructed. We define three classes for the activities: (1)




(a) Talking head (b) Writing hand

(c) Slide show (d) Writing hand

Figure 2. Types of frames in lecture video. (a) is a non-content frame, (b), (c) and (d) are content frames.

talking head, (2) writing hand and (3) slide show. The first one is comprised of non-content frames and the other two, of content frames. Motion in the first class is more, compared to that in the second which in turn, is greater than that in the third. Hence the energy of the temporal derivative in intensity space can be used as a relevant feature, which is given by

$$x_1(t) = \sum_{m=1}^{M} \sum_{n=1}^{N} (F_t(m,n) - F_{t-1}(m,n))^2 \qquad (2)$$

where $F_t$ is the pixel intensity values of the current frame and $F_{t-1}$ is those of the previous frame, $M$ is the number of rows and $N$ is the number of columns in the video frame.

The gray-level histogram $h_t(i)$ gives the distribution of image pixels over different intensity values. The histogram will be very sparse for the slide show class and moderately sparse for the writing hand and dense for a talking head. Hence the entropy of the histogram [14, 15] can be treated as another good feature for the detection of these activities, which is given by

$$x_2(t) = -\sum_{i=0}^{255} h_t(i) log(h_t(i)). \qquad (3)$$

We use an HMM with the Gaussian mixture model(GMM) assumption, with two number of states. The Baum-Welch algorithm is used for estimating the parameters [10]. In this, the parameters of HMM are initially set with random values for each class and then we employ the expectation-maximization(EM) algorithm [16], which in a few iterations gives well-tuned parameters for each class.

In the recognition phase also, same features are extracted from each image of the frame sequence of the test data. These features are compared against the models

for each class of instructional activity . For a classifier of $C$ categories, we choose the model which best matches the observations from $C$ HMMs and the class that gives the highest likelihood score is declared as the recognized class. This yields very good results, even on training with one lecture video sequence and testing it on another as the scene complexity is quite low.

### *B. Visual quality assessment of non-content frames*

The talking head activity, which is a non-content one, is to be represented by an appropriate video clip on media re-creation. The selection of this is done using an objective, no-reference perceptual quality assessment method [7, 8]. In this, blurring and blocking effect are used as the most significant artifacts, generated during the image compression process. Let the frame be $F(m,n)$ for $m \in [1,M]$ and $n \in [1,N]$. A difference signal along each horizontal line is computed as :

$$d_h(m,n) = F(m,n+1) - F(m,n), n \in [1,N-1]. \quad (4)$$

The blockiness is estimated as the average differences across block boundaries :

$$B_h = \frac{1}{M[(N/8)-1]} \sum_{i=1}^{M} \sum_{j=1}^{(N/8)-1} \mid d_h(i,8j) \mid . \quad (5)$$

Since blurring causes the reduction of signal activity, combining the blockiness and activity measure is more useful to deal with the relative blur in the frame. The activity is measured using two factors, the first one being the average absolute difference between in-block image samples

$$A_h = \frac{1}{7}\left( \frac{8}{M[N-1]} \sum_{i=1}^{M} \sum_{j=1}^{N-1} \mid d_h(i,j) \mid - B_h \right) \quad (6)$$

and the second one being the zero-crossing(ZC) rate. We define for $n \in [1,N-2]$

$$z_h(m,n) = \begin{cases} 1; & for\ horizontal\ ZC\ at\ d_h(m,n) \\ 0; & otherwise. \end{cases}$$

The horizontal ZC rate is estimated as :

$$Z_h = \frac{1}{M[N-2]} \sum_{i=1}^{M} \sum_{j=1}^{N-2} z_h(i,j). \quad (7)$$

Then we calculate the vertical features $B_v$, $A_v$ and $Z_v$ by a similar procedure and finally, the combined features as:

$$B = \frac{B_h + B_v}{2}, A = \frac{A_h + A_v}{2}, Z = \frac{Z_h + Z_v}{2}. \quad (8)$$

The quality measure is given by [7, 8]

$$QS_N = a + bB^{c_1}A^{c_2}Z^{c_3} \quad (9)$$

where $a, b, c_1, c_2$ and $c_3$ are the parameters usually estimated using subjective test data. We use the same values as suggested in [7, 8]. Using this, quality scores are assigned to the frames from the talking head portion of the lecture video sequence. The frame positions with locally high quality scores are noted for the selection of video clips for highlight creation.

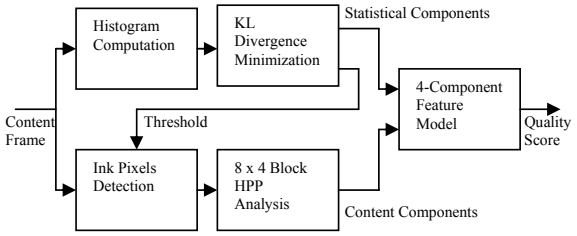



Figure 3. Block schematic for the proposed method of quality assessment of content frames.

### *C. Visual quality assessment of content frames*

For the selection of video clips for the content segments, we propose a four component based feature for visual quality assessment which is depicted in Fig. 3. First, the content frame is converted into gray scale and the histogram $h(i)$ is computed. It is modeled by a bimodal(one corresponding to ink pixels and other to the rest) GMM, whose *pdf* is given by :

$$p(i) = \frac{\epsilon}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{i-\mu_1}{\sigma_1}\right)^2} + \frac{1-\epsilon}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{i-\mu_2}{\sigma_2}\right)^2} \quad (10)$$

where $i$ is the intensity level, $\epsilon$ is the proportion of the mixture, $\mu_1$ is the foreground mean, $\mu_2$ is the background mean, $\sigma_1^2$ is the foreground variance and $\sigma_2^2$ is the background variance. In order to compute these parameters [17], we minimize the Kullback Leibler(KL) divergence $J$ from the observed histogram $h(i)$ to the unknown mixture distribution, $p(i)$. $J$ is given by

$$J = \sum_{i=0}^{255} h(i) log\left[\frac{h(i)}{p(i)}\right]. \quad (11)$$

Since the numerator term $h(i)$ does not depend on the unknown parameters, the minimization is equivalent to minimizing the information measure $Q$, where

$$Q = -\sum_{i=0}^{255} h(i) log[p(i)]. \quad (12)$$

To carry out the minimization, we assume that the modes are well separated. If $T$ is the threshold which separates the two modes, we have

$$p(i) \approx \begin{cases} \frac{\epsilon}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{i-\mu_1}{\sigma_1}\right)^2}; & 0 \le i \le T \\ \frac{1-\epsilon}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{i-\mu_2}{\sigma_2}\right)^2}; & T < i \le 255 \end{cases}$$

Now

$$Q(T) = -\sum_{i=0}^{T} h(i) log\left(\frac{\epsilon}{\sqrt{2\pi}\,\sigma_1}\right) e^{-\frac{1}{2}\left(\frac{i-\mu_1}{\sigma_1}\right)^2} - \sum_{i=T+1}^{255} h(i) log\left(\frac{1-\epsilon}{\sqrt{2\pi}\,\sigma_2}\right) e^{-\frac{1}{2}\left(\frac{i-\mu_2}{\sigma_2}\right)^2}. \quad (13)$$

Minimizing $Q(T)$ with respect to $\{\mu_1, \sigma_1, \mu_2, \sigma_2, \epsilon, T\}$ gives the statistical parameters along with the optimum threshold for ink pixel detection of content frames. The

mean and variance of the foreground and background can be used to define the quality of content frame. Hence we define the following terms which contribute to quality:

$$MeanTerm = C = \frac{\mid \mu_2 - \mu_1 \mid}{255}. \quad (14)$$

$$SigmaTerm = S = \frac{1}{\sigma_1 + \sigma_2}. \quad (15)$$

These terms take care of the statistical aspect only and in order to include the spatial arrangement of the content, we now use the HPP based method [18, 19] to derive a pair of features.

The ink pixel detected frame is partitioned into $8 \times 4$ equal blocks. This division into blocks will take care of the extent to which the frame is filled 'uniformly' by ink pixels across the entire document page. For each block, the HPP is constructed. It is obtained by summing the number of ink pixels in each row of the block along horizontal direction. It is an array $F_b$ of $K$ elements, which are normalized in [0,1]. Now the energy of each HPP is calculated by

$$E_b = \frac{1}{K} \sum_{m=1}^{K} \mid F_b(m) - \hat{F}_b \mid^2; b = 1, 2, ..., 32. \quad (16)$$

where $E_b$ is the energy of the HPP of the block $b$, $F_b(m)$ is the HPP of the block $b$ and $\hat{F}_b$ is the average value of HPP of block $b$. Low value of energy indicates predominantly black or white patches while a high value indicates well-written block. Now these are added to get the energy of a frame

$$E = \sum_{b=1}^{8\times4} E_b \quad (17)$$

which is another measure of content quality.

The average value of the HPP of individual blocks, $\hat{F}_b$ can be used as another cue to diminish the quality score of those frames with patches of occluding hand or picture-in-picture. For this, we perform an inter-block comparison to calculate the difference

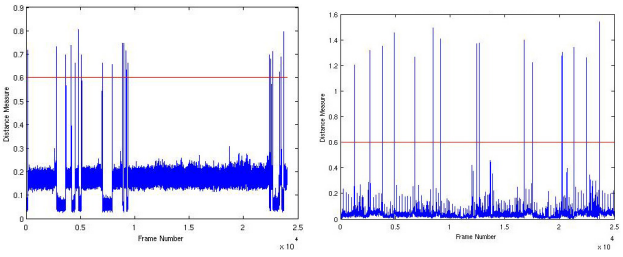
$$d = \max_b \hat{F}_b - \min_b \hat{F}_b.$$

If $d$ is small, all blocks are 'clean' and free from patches. If $d$ is high, there is the presence of a patch or occluding objects. Hence $G = 1/d$ can be used as another measure of quality.

We have defined four terms for quality assessment of content frames. Now a weighted sum of these individual scores are taken to get the final quality score.

$$QS_C = \alpha_1 E + \alpha_2 S + \alpha_3 C + \alpha_4 G. \quad (18)$$

These weights $\alpha_i$s are estimated using subjective test data of content frames. Using this, the quality assessment of content frames are done and the frame positions with high quality scores are noted for the selection of video clips for highlight creation.




(a) (b)

Figure 4. Plot of histogram difference measure for a lecture video containing (a) printed slides and hand written portions, (b) writing hand segments. The horizontal line indicates the threshold for shot detection.

### D. Media re-creation for capsule preparation

The locations for high quality frames give the temporal marking around which the desired video clips are to be selected, along with the audio. On media re-creation, we select those frames corresponding to $\pm 5$ seconds around these high quality frame instants provided there is no shot change within this period, to produce the highlight. The choice of a 10 sec long window is meant to convey the instructional content to the viewer for a single highlight. We select such 10 sec windows around each of the prominent local peaks in the quality measure. On subjective evaluation, we found that an appropriate temporal proportion of the recognized classes during capsule preparation is 1:3 if there are only two classes, namely talking head and writing hand. If talking head, writing hand and slide show classes all occur, then an appropriate ratio was found to be 1:2:2.

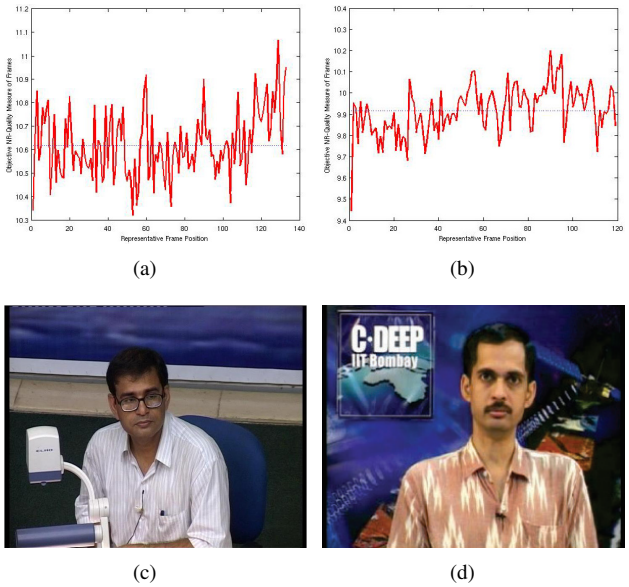
As a matter of fact, the first clip shown is the talking head to show the instructor, then only the other two classes in suitable proportions. The audio stream is kept in perfect synchronism with the visual data. Typically a 1 hour lecture may contain 10 to 12 video clips, each of about 10 second duration, which yields a lecture video capsule of approximately 2 minutes. Media re-creation is done using FLASH [20].

## III. Results

We worked on lecture videos of different instructors, each of 1 hour duration. These videos contain either handwritten slides or computer generated slides or both.

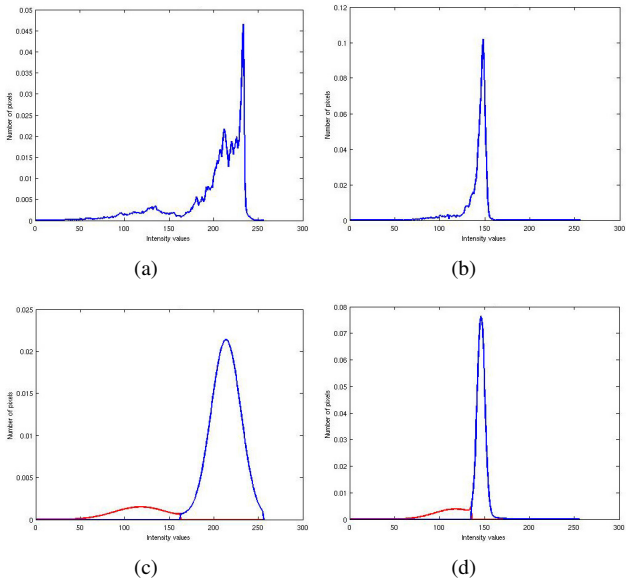
The first phase of temporal segmentation has effectively identified the scene breaks in the lecture video sequence. The plots of the histogram difference measure, obtained from two videos are shown in Fig. 4 (a) and (b). The spikes in these plots represent the possible scene breaks. These are detected against a threshold to effect the segmentation. As it can be seen, this histogram difference measure works well, as there is very little camera and object movement.

In the case of activity detection, the training of HMM was done with the three classes of video activities as already mentioned, by which it could effectively classify a given test sequence. Referring to Fig. 2 again, (a), (b) and (c) give the representative frames from talking head, writing hand and slide show sequences, respectively used



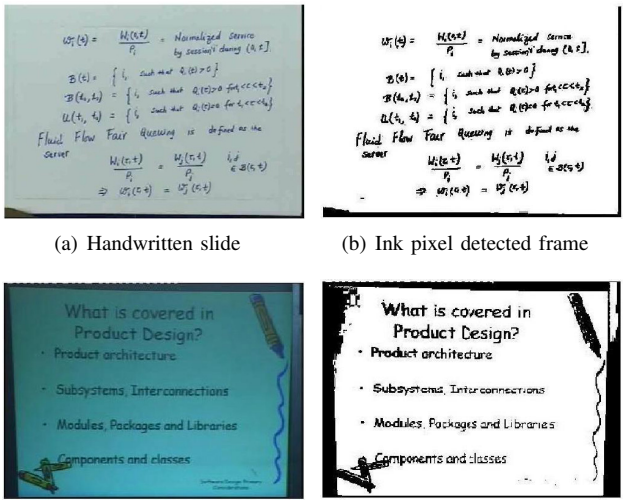

(a) (b) (c) (d)

Figure 5. Results of no-reference quality assessment of talking head frames. (a) & (b) are the plots of the quality measure for different frame sequences, (c) & (d) are the corresponding selected frames with the best quality.




(a) (b) (c) (d)

Figure 6. Plot of observed and estimated histograms for content frames (a) & (b) are observed histograms, (c) & (d) are the corresponding estimated histograms with optimum threshold.




(a) Handwritten slide (b) Ink pixel detected frame

(c) Electronic slide (d) Ink pixel detected frame

Figure 7. Content frames and their corresponding ink pixel detected frames. (b) and (d) are the detected ink pixels.

for training and (d) is a sample frame from the correctly detected activity of writing hand.

The results of no-reference objective measure based quality assessment for non-content frames are shown in Fig. 5. We notice that frame 5128 offers the best quality of the talking head for the first frame sequence and that the frame 3414 does it for the second sequence. The parameters of the quality score evaluation, as given in equation (9), obtained with all test images are $a = 245.9, b = 261.9, c_1 = 0.0240, c_2 = 0.0160, c_3 = 0.0064$.

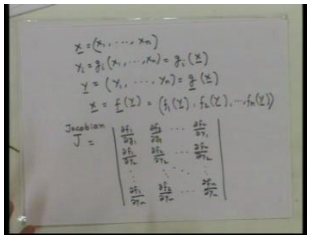
In the case of content frames, the quality assessment process starts with the bimodal GMM assumption of the histogram of the frame and a constrained optimization to yield the statistical parameters along with the optimum threshold. Results of the KL divergence minimization to yield an estimate of the bimodal GMM parameters is depicted in Fig. 6. The optimization started with an initial condition of $x_0 = [\mu_1, \sigma_1, \mu_2, \sigma_2, \epsilon, T] = [100, 20, 200, 10, 0.1, 150]$ and a typical obtained output is $[39.56, 33.83, 150.63, 5.30, 0.21, 132]$. The last element is the optimum threshold used for ink pixel detection, before HPP analysis. Results of this are shown in Fig. 7.

The $8 \times 4$ block based HPP analysis is found to be effective in quantifying the spatial distribution of content and hence the quality of a frame. Two sets of results of this are shown in Fig. 8. Note that a white patch(block) in the frame results in an average HPP value of 1 and a resultant HPP energy of 0. The quality assessment of content frames based on the four component feature model is performed with the following weights: $\alpha_1 = 1, \alpha_2 = 50, \alpha_3 = 4, \alpha_4 = 0.4$. The results of the quality assessment of handwritten content frames are shown in Fig. 9 and those of slide show content frames are shown in Fig. 10. We found the results to be quite satisfactory for capsule preparation.

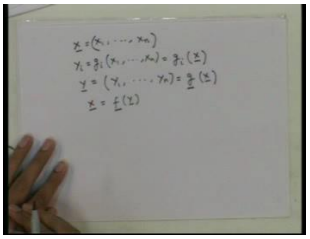
The media re-creation to produce lecture video capsule is done using FLASH. For this, 250 frames, around the high quality frame instants are selected as highlights. Our algorithm, worked on the lecture video of 1 hour duration yielded a capsule of 2 minutes, with suitable proportion of instructional activity highlights.

## IV. Conclusion

We developed an efficient algorithm for automatic capsule preparation of instructional video. As demonstrated, visual quality based highlight extraction is performed differently on separate classes of instructional activity and the selected video clips are merged in suitable proportion to create a compact video capsule. This video capsule



(a) A high quality content frame



(b) A low quality content frame

| | | | |
|---|---|---|---|
| 0.92 | 0.89 | 0.93 | 1 |
| 0.93 | 0.86 | 0.8 | 0.99 |
| 0.94 | 0.89 | 0.81 | 0.9 |
| 0.96 | 0.86 | 0.8 | 0.85 |
| 0.83 | 0.75 | 0.81 | 0.96 |
| 0.95 | 0.76 | 0.75 | 0.97 |
| 1 | 0.89 | 0.93 | 0.97 |
| 0.92 | 0.8 | 0.75 | 0.95 |

(c) $8 \times 4$ average HPP values

| | | | |
|---|---|---|---|
| 0.9 | 0.85 | 0.89 | 1 |
| 0.91 | 0.82 | 0.73 | 0.99 |
| 0.94 | 0.82 | 0.84 | 0.98 |
| 0.95 | 0.86 | 1 | 1 |
| 0.95 | 1 | 1 | 1 |
| 0.55 | 1 | 1 | 1 |
| 0.35 | 0.95 | 1 | 1 |
| 0.1 | 0.61 | 1 | 1 |

(d) $8 \times 4$ average HPP values

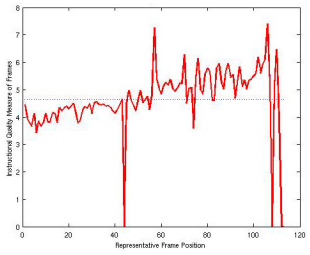
| | | | |
|---|---|---|---|
| 0.05 | 0.09 | 0.07 | 0 |
| 0.06 | 0.1 | 0.12 | 0.02 |
| 0.05 | 0.06 | 0.14 | 0.1 |
| 0.04 | 0.13 | 0.15 | 0.17 |
| 0.14 | 0.1 | 0.08 | 0.01 |
| 0.07 | 0.06 | 0.08 | 0.01 |
| 0 | 0.06 | 0.04 | 0.01 |
| 0.06 | 0.07 | 0.08 | 0.03 |

(e) $8 \times 4$ block HPP energies

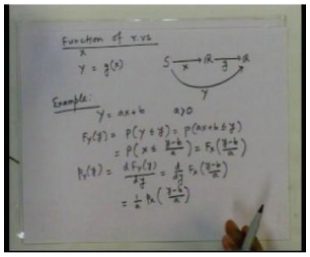
| | | | |
|---|---|---|---|
| 0.07 | 0.12 | 0.12 | 0 |
| 0.08 | 0.14 | 0.17 | 0.02 |
| 0.06 | 0.1 | 0.16 | 0.02 |
| 0.04 | 0.13 | 0 | 0 |
| 0.08 | 0 | 0 | 0 |
| 0.03 | 0 | 0 | 0 |
| 0.01 | 0.06 | 0 | 0 |
| 0.12 | 0.12 | 0 | 0 |

(f) $8 \times 4$ block HPP energies

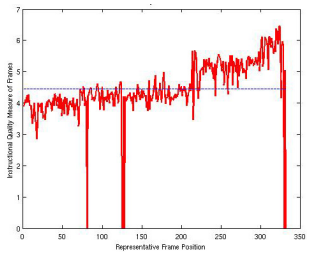
Figure 8. Results of $8 \times 4$ block HPP analysis. (a), (c) & (e) are one set of result for a high quality content frame and (b), (d) & (f) are another set for a low quality content frame



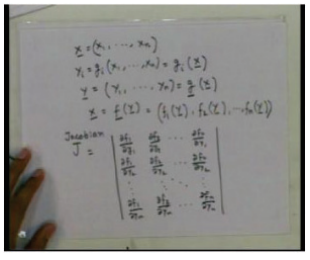

(a)



(b)




(c)



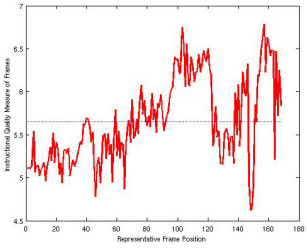
(d)

Figure 9. Results of quality assessment of hand written content frames. (a) plot of quality score for one segment, (b) corresponding best quality frame, (c) plot of quality score for another segment and (d) corresponding best quality frame.

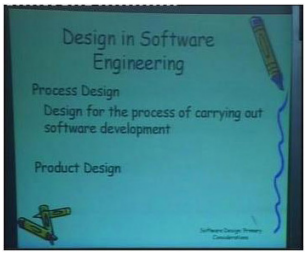
helps the users for a fast preview of the entire course lecture.

## REFERENCES

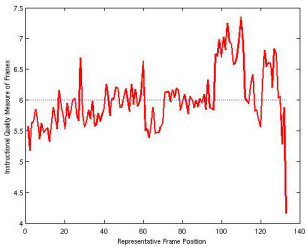

[1] T. Liu and J. R. Kender, "Lecture videos for e-learning: Current researches and challenges," *Proc. of the IEEE Sixth International Symposium on Multimedia Software Engineering*, 2004.

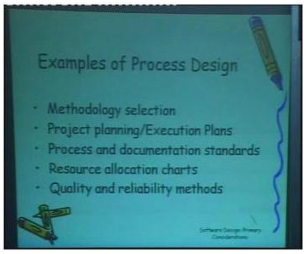

(a)




(b)




(c)




(d)

Figure 10. Results of quality assessment of slide show content frames. (a) plot of quality score for one segment, (b) corresponding best quality slide, (c) plot of quality score for another segment and (d) corresponding best quality slide.


[2] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short : Dynamic video synopsis," *Proc. of the 2006 IEEE Computer Society Conference on CVPR*, vol. 1, pp. 435–441, June 2006.

[3] S. Dagtas and M. Abdel-Mottaleb, "Extraction of tv highlights using multimedia features," *Proc. of fourth IEEE workshop on Multimedia signal processing*, 2001.

[4] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for tv baseball programs," *Eighth ACM International Conference on Multimedia*, pp. 105–115, 2000.

[5] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hmm," *Proc. of the IEEE International Conference on Image Processing*, 2002.

[6] T. Syeda-Mahmood and D. Ponceleon, "Learning video browsing behaviour and its application in the generation of video previews," *ACM Multimedia*, pp. 119–128, October 2001.

[7] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," *Proc. IEEE International Conference on Image Processing*, pp. 477–480, September 2002.

[8] H. R. Sheikh, A. C. Bovik, and L. K. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg 2000," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 1918–1927, December 2005.

[9] L. R. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, January 1986.

[10] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, pp. 257–286, 1989.

[11] N. Robertson and I. Reid, "A general method for human activity recognition in video," *Computer Vision and Image Understanding, Elsevier*, vol. 104, no. 2, pp. 232–248, November 2006.

[12] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, August 2000.

[13] F. Niu and M. Abdel-Mottaleb, "Hmm-based segmentation and recognition of human activities from video sequences," *IEEE International conference on Multimedia and Expo*, July 2005.

[14] N. R. Pal and S. K. Pal, "Entropy: A new definition and its applications," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, no. 5, pp. 1260–1270, October 1991.

[15] ——, "Object-background segmentation using new definition of

entropy," *IEEE Proceedings*, vol. 136, no. 4, pp. 284–295, July 1989.

[16] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, November 1996.

[17] J. Kittler and J. Illingworth, "On thresold selection using clustering criteria," *IEEE Transactions on SMC*, vol. 15, pp. 652–655, 1985.

[18] S.-W. Lee and D.-S. Ryu, "Parameter-free document layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1240–1256, November 2001.

[19] S. N. Srihari and V. Govindaraju, "Analysis of textual images using the hough transform," *Machine Vision and Applications*, vol. 2, no. 3, pp. 142–153, June 1989.

[20] Web-document, "Echoecho.com tutorials : Flash tutorial," *http://www.echoecho.com/flash.htm*.