For video lecture transmission, less is more: Analysis of Image Cropping as a cost savings technique

by

Sai Gollapudi, Niyati Mishra

in

APCHI ’13 Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction

Report No: IIIT/TR/2013/-1

Centre for Software Engineering Research Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
September 2013
For video lecture transmission, less is more: Analysis of Image Cropping as a cost savings technique

Gollapudi V. R. J. Sai Prasad  
IIIT-Hyderabad  
Gachibowli, Hyderabad, AP, 500032, India  
saigollapudi1@gmail.com

Niyati Mishra  
IIIT-Hyderabad  
Gachibowli, Hyderabad, AP, 500032, India  
niyati.mishra88@gmail.com

ABSTRACT
This paper explores the implications of socio-cultural constraints on the design of a video deployment system. The research is motivated by real world difficulties faced by students in downloading large course-content videos. For instance, online lecture videos may rarely be downloaded by students because of their institution’s bandwidth constraints. One simple method to lessen the bandwidth burden is to have a deployment system that crops out non-essential portions of the transmitted video. This paper explores how cropping may adversely impact a student’s impressions and understanding of the transmitted content. A two part, between-subject experiment was conducted. The results: when cropping was applied, the file sizes decreased, and the message clarity increased by 12%. Moreover, cropping did not significantly lessen the personality judgment the students were making of the instructor. So, post-processing of videos for transmission, by cropping non-essential content, does show promise, while not compromising content or impressions.

Author Keywords
video cropping; NPTEL videos; cropping gesture; perceptual engineering

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION
Having thousands of hours of course content videos (like the NPTEL videos [15], generated under the Ministry of HRD, Govt. of India) is a great digital asset. However, its use can be, and also needs to be, further maximized and fully leveraged. We feel that for tier 2 and tier 3 educational institutions and students there are practical problems that forbid them to easily incorporate this existing content into their academic lives. Two problems that we perceive are: One, poor infrastructure which forbids quick downloading of large quantities of video content. And two, lack of access to enough computer systems that makes it difficult for interested students to watch hours and hours of video on dedicated computer systems. So, bottom line: How to better serve the institutes that lack good download capability? And, how to better serve the students who lack access to dedicated computer systems for viewing these downloaded videos? Can post processing of these already completed NPTEL videos help in resolving these socio-cultural constraints? In this research, our aim was to explore ways to reduce the footprint of a (NPTEL like course-content) video, without compromising on either the resolution or on the technical content. Such smaller sized videos, we felt, could reduce the transmission bandwidth requirement, and also give an option of making the videos viewable on portable and lower cost devices, like a ubiquitous mobile phone. Therefore, the specific research question that we chose to explore here was: What impact does cropping of non-essential video content have on student’s understanding and impressions of an educational, lecture style, course content video?

Here, the focus of research was not directly technical in that bandwidth optimization techniques [8, 6], or video compression techniques [9] were not researched. For the sake of bandwidth efficiencies, neither the video content was transformed [14, 2], nor the transmission protocol modified [16]. Instead, the research focused on the human (student), system (video) interface and usage. And in that relationship between the student and the video, the focus was more on study of the implication of removal of certain portions of video on student perceptions and also on bandwidth reduction. It is expected that the findings from this study could help inform the design of video deployment systems.

Logic behind cropping technique
Assuming that a typical video mediated system currently transmits ‘all that the camera sees’, we question if, at the receiving end, a human (perceptual system) does indeed require all that which is being visually transmitted. We suggest that the raw data that is being transmitted consists of both, ‘essential’, as well as, ‘non-essential’ information. We hypothesize that by cropping the ‘non-essential’ information from the transmission, we not only reduce the bandwidth usage, but also not suffer any significant message loss. In the context of educational technologies, remote and digital classrooms often rely on pre-recorded video lectures to deliver course content [15]. In contrast to entertainment videos, these video lectures are often taped in front of a content-irrelevant still background (usually a blackboard or a university logo as the backdrop), with no props, and often lack in emotional substance.
From a cinematographic point of view, the video lectures are often midlevel close-up shots of a talking head with little or no variation in the visual shot.

Assuming that general human dialog consists of both verbal and non-verbal cues, then one can categorize the spoken content of these video lectures as the verbals, and the gesture information, the background stills, and the instructor’s body appearance visuals as the non-verbals. For course content delivery, it is assumed that the verbals are essential and play a larger role than the non-verbals; thus, providing an opportunity to crop them out for saving of bandwidth. But such bandwidth savings due to cropping may come at a cost. For instance, these non-verbals may be key to comprehension of Emotion, Interpersonal Attitudes, Self Presentation [1] of the instructor etc.

What happens to the human perceptual system when only portions of pre-recorded course video are sent? What if the background is fully cropped out as non-essential data? What happens if the instructor’s face is the only thing that is shown, and no hand or body gestures are shown? Will the student get confused for not having received the ‘non essential’ data? Will the student still be able to make out the gist of the message transmitted by the instructor? What specific communication is lost when instructor’s hand and body gestures are disabled? These were the questions that motivated this research.

This study of human perceptual processing revolved around two hypotheses: (h1) not transmitting instructor’s background visual, or instructor’s hand or body gesture information, does not significantly impede on the student’s understanding of the transmitted message. And, (h2) the student’s subjective assessment of the personality of the instructor will be less, when the student is visually deprived of having instructor’s hand and body gesture (non-verbal) information.

RELATED WORK
With the advent of mobile devices, watching movies and videos on smaller screens has become a common practice. The process of reformatting the original video size to suit smaller screens [4] and tiny devices [17, 7] has already been studied. Video Retargeting [13] is one of the terms used to describe this type of screen readjustment activity. To meet today’s demand of multiple devices, a variety of screen sizes and high quantity of video content, Video Retargeting is having to be automated. And, in automation, the key questions that researchers are having to address are: what part to crop and what not to crop? what will keep and ensure that the original content and intent are in tact [17, 4, 12]?

In our research, the focus is not primarily on the video resizing issue. Instead, the goal is to explore how video resizing can be used for saving on bandwidth. This approach has more to do with reducing the file-size footprint of the video than with resizing. When insights from Video Retargeting work are applied to our earmarked genre of Lecture Videos, we notice few things which makes the retargeting easier to automate. One, the Region of Interest (ROI) [3] for video lectures – especially for those with only talking heads – is typically in the center and is less dynamic. That is, less pre-processing, less focus on finding out what is important in a scene, and more static cropping can be applied. Two, we are finding that background content is sometimes not supportive to the content, and at times, is potentially distracting to the viewer.

Our contribution to the bandwidth conscious community is that we can potentially apply video retargeting insights to reduce video transmission bandwidth. Our contribution to the video retargeting community is that for certain genre of educational videos, we can use simple static cropping to increase the significance of the content in the resized output to the end user. From our point of view, both are non-intuitive contributions.

METHOD
To study the implications of cropping as a bandwidth treatment, a three piece experiment was setup: One of the important pieces was indeed the video transmission portion; but the other two components were the video generation component in the front, and the video consumption component at the receiving end. In this setup the impact of treatment to bandwidth is studied by comparing the generated video-message with that of the received video-message.

Experiment
To test our two hypotheses we conducted a two part, between-subjects, mixed format experiment. The intent of the first phase called herein as P1 was to generate video output that was naturally gesture-rich. This represented the pre-recorded course lecture videos that were typically transmitted in remote or digital classrooms.

The intent of the second phase called herein as P2 was to, one, capture the comprehension levels for various formats of input stimuli, and two, to access the emotional or social nature of the virtual connection between the receiver and the sender. In the educational context, the participants represented the students who typically viewed video lectures on a remote or a digital classroom.

Dividing the experiment into two parts (see Figure 1) P1 and P2 enables the separation of the creation (P1) from the consumption (P2) of the video-stimulus. It also provides a great opportunity to compare and contrast the various transmission formats and their impact on participant (i.e. student) comprehension; And to also study the implication of the transmission format on the social level personality assessment of the instructor by the student.

Video Creation Activity (P1)
The objective behind P1 was to generate afresh some video content that suitably represented (and also substituted for) a pre-recorded video lecture. The reason for creating video afresh was that in P2, when the comprehension of the transmitted video was assessed, having a small, complete and to-the-point video helped. Moreover, with a stimuli inspired video, there was also a better opportunity to judge comprehension. That is, if a receiver was able to trace the fresh video back to the stimuli which inspired it, one can conclude
that the gist of the video message got through to its audience; which was what was essentially to be tested.

By having control on the video generation, in P1 the length of the video could be minimized (for example, to 2-3 minutes), and also the intended message for transmission could be made clear. In contrast, with a pre-recorded course lecture video, the length of the video varied (sometimes from 15-20 minutes to 1+ hour) and also the real intent of a specific lecture was often debatable.

In the capturing of the video for P1 the video content was sought to be gesture-rich and also natural in its conversational flow. For validating (h2) this P1 video also needed to contain content that was subtly loaded with personal, social or emotional material, such that it could be detected by receivers in P2. To instigate such output from the participants a Picture Description Task (PDT) and a paired Picture Identification Task (PIT) were used. The exact Projective Test in psychology which incorporates this is called the Thematic Apperception Test or TAT for short [5]. Though ours is not a true TAT, it is nevertheless inspired by it. Henceforth, for convenience, we refer to this task in the experiment as PDT. Similar to TAT, our participants too were shown ambiguous pictures and were asked to respond to them on camera. Their animated story telling about the stimuli was recorded as the message from the sender, to be transmitted to the receiver at the other end of this virtual, video mediated transmitting system.

**Video Transmission**

In this experiment, there was no explicit transmission of video. Neither was there any explicit bandwidth measurement made. Here transmission is only assumed to happen; it is implicit. When a type of video is shown to participants in P2, the assumption is that such a transmission must have taken place from sender to receiver. Likewise, when a video is cropped and only a portion of the original is shown, the claim is that the cropped portion of the video is the saving on the bandwidth and transmission cost.

In reality various codecs and other transmission protocol could be applied to the data. Our bandwidth saving proposal is prior to and complementary to such data transmission activity.

For calculating bandwidth usage and savings, we assume raw numbers (before a codec) and show the savings at the bit/byte rates, achieved before any data transmission activity.

**Video Consumption Activity (P2)**

Objective of P2 was to measure message comprehension, quality of media and subjective assessment of the sender of message. These were to be taken not only for the raw video that got generated in P1, but also for the other transmission treatments that got generated from P1.

Overall there were three video formats that is, three treatments that got generated for use in P2. They were one, the raw video, two, a cropped video, and three, just the audio. Here the original raw video was for control. The cropped video was the actual bandwidth saving technique that was being tested. Having a third format audio only was to generate a lower bound measure for the test.

In the experiment, if P1 represents an instructor putting together a course content video, then P2 represents the community of students that view and learn from that video. So, while P1 features very few participants, then P2, in contrast, will involve multiple participants.

In the experiment, bandwidth savings was to be demonstrated by measuring (h1) comprehension of message by students, and showing that even by transmitting lesser information, the transmission system was still able to convey the same content or message to the students.

To this end, the P2 participants were divided into three groups (Table 1), with each receiving one of the three transmission versions of the video.

In our between-the-subjects and mixed design, the participants of group 1 received the original raw video; group 2 members received the cropped video; and the last group 3 members received the audio version of the P1 video.

A message-comprehension test and a media-clarity survey (or channel quality test) together highlight the viability of using each of the transmission treatments as an option for saving

<table>
<thead>
<tr>
<th>TAT photo stimuli</th>
<th>full, raw video</th>
<th>video cropped to passport size</th>
<th>audio &amp; no video</th>
</tr>
</thead>
<tbody>
<tr>
<td>photo 1 / story 1</td>
<td>group 1</td>
<td>group 2</td>
<td>group 3</td>
</tr>
<tr>
<td>photo 1 / story 2</td>
<td>group 1</td>
<td>group 2</td>
<td>group 3</td>
</tr>
<tr>
<td>photo 1 / story 3</td>
<td>group 1</td>
<td>group 2</td>
<td>group 3</td>
</tr>
<tr>
<td>photo 1 / story 4</td>
<td>group 1</td>
<td>group 2</td>
<td>group 3</td>
</tr>
<tr>
<td>photo 1 / story 5</td>
<td>group 1</td>
<td>group 2</td>
<td>group 3</td>
</tr>
<tr>
<td>photo 1 / story 6</td>
<td>group 1</td>
<td>group 2</td>
<td>group 3</td>
</tr>
</tbody>
</table>

Table 1. Grouping of Participants in P2.
on the bandwidth. For any of the transmission formats, if the values from these two tests were poor, then it can be inferred that that particular format is not viable for bandwidth savings. Instead, if the results of these two tests proved to be good, then the format is acceptable as a way to reduce the transmission bandwidth usage. This helps in establishing (h1).

For establishing (h2) quantitative assessment data for the instructor had to be captured. An existing Personality Inventory Index called HEXACO-60 was used [10, 11] for this purpose.

Participants
P1 had 3 participants who were all male and university students. From the shortlisted three, after taping and observing the quality of the video, only 1 of the 3 was selected for use in P2. The criteria for the selection of the final video was subjective, and, was based on the richness of the narrative, the potential for subtle emotional content, and the quantity of gestures.

For P2 there were 30 undergraduate university students, of which 60% were males. The average age of participants was 22.5 years. All the participants were from India and were noticeably familiar with the Indian street environment (which was the theme of the P1 PDT photo stimuli).

Stimulus
For P1, six photographs were provided for the PDT test. These photographs were downloaded from Google image search, and were each shown on their own individual Microsoft Power Point slide. The choice of the photographs was arbitrary. However, in the selection it was ensured that the photographs had some level of subject ambiguity so that it was capable of triggering a creative story.

The first photograph was that of a lady in a Sari standing still. Similarly, for the remaining photos our distractors really meant that the student got the transmitted message. Thus a participant accurately identifying a picture from a set of distractors was approximates of the originals, with the exception of one clue being different from each and the rest. This maintained enough of a distraction, while still giving a chance for those who got the gist of the message to identify themselves.

The second photograph was a black and white photograph showing children playing in the rain. The third photograph was the theme of the P1 PDT photo stimuli). For another distractor photograph we incorporated a picture of a lady in a Sari standing still in a busy marketplace (Figure 2). The picture itself was ambiguous about what she was doing or not doing in the busy bazaar. The photograph does not conclusively offer any suggestions about her emotional state or actions. This ambiguity is what was planned for triggering a narrative by the P1 participant.

The second photograph was a black and white photograph showing children playing in the rain. The third photograph showed two young ladies cheering loudly into the ear of a neighboring security guard (apparently at a sporting event). A photo of a street vendor selling coconuts, a picture of an auto rickshaw in motion, and a photo of a Indian street scene were the subjects of our fourth, fifth and sixth photographs. The stimulus for P2 came from the processing of P1 generated video. The first version was the original raw version. In this case there was no tampering of the original P1 output. This video was being transmitted at 24 frames/sec, with each frame consisting of 1080x720 pixels. True Color with bits per pixel rate of 24 bits per pixel were used. This resulted in 1080x720x24 bits/frame.

In the second case, the original video was cropped to 'passport photo' size. (See Figure 3) By doing this a 1080x720 pixel per frame photo was reduced to a 150x150 pixel per frame photo. The cropped quantity is 7,55,100 bits per frame.

In audio only format, no video was transmitted. Audacity was the audio editor that was used for taking out extraneous sounds like researcher’s voice, door sound etc.

In P2, after viewing of the videos, the participants were shown a series of 24 photographs, as a single collage, on one Microsoft Power Point slide (Figure 4). The grid was a 6 row by 4 column table, showcasing 3 distractors for every one of the 6 initial PDT photo stimuli.

The distractor images were also Google search downloads of images that closely corresponded to the original 6 PDT stimuli. For instance for one Photo 1 distractor, we used a photo of another sari clad women, also taken in a bazaar. The difference here was that this woman was walking in the bazaar, instead of standing still (as was the case in the original PDT photo stimuli picture).

For another distractor photograph we incorporated a picture of a lady in a bazaar purchasing, but this one was not clad in a sari. In a third distractor we showed not a woman but a little girl in a sari. Similarly, for the remaining photos our distractors were approximates of the originals, with the exception of one clue being different from each and the rest. This maintained enough of a distraction, while still giving a chance for those who got the gist of the message to identify themselves.

The stimulus was such that if the student received any portion of the transmitted message, then he or she would not be able to pick the correct picture from the collage. Thus a participant accurately identifying a picture from a set of distractors really meant that the student got the transmitted message.

Procedure
In P1, after registration, the participants were asked to view a photograph. This was a PDT photo stimulus. Here the participant was instructed to reflect upon the photograph and extemporize compose his or her own story justifying the content of the photograph. This narrative was videotaped and stored. For P1 PDT test, 6 photos were shown and 6 corresponding narratives were recorded and stored as story 1, 2-6; or, narratives 1, 2-6.
These 6 captured videos were further processed in three ways: For one, the raw video was kept as is; for two, the raw video was cropped to show bandwidth usage savings; and for three, only the audio track was kept, and the video track was fully removed. This step was done in the background, before embarking on P2. Processing 3 different versions of the original 6 narratives resulted in 18 final samples for use in P2. At the beginning of P2, the participants were arbitrarily divided into three groups. Each group was shown all the 6 narratives, in one of the three formats. For instance, group one would have seen all 6 narratives of the original raw video format. Group two was perhaps shown all 6 narratives in the cropped format only. And group three, might have heard all 6 narratives presented in audio format only.

After viewing each of the 6 narratives, each group was asked to select one photograph from a collage of 24 distractors that would best fit the just narrated story. Each participant’s selection of 6 photographs was then recorded. This was to test the participant’s message comprehension.

Post this test, the participants were asked to answer three questions about the media and the quality of the channel. This was a multiple choice Likert 5 measure survey. Here the ambient noise, and potential confusion factor in the transmitted video signal was reported and recorded.

The Likert measures used for the first two questions were very poor (-2), poor (-1), neutral (0), good (+1), very good (+2). For the last of the three questions, which related to ambient noise over signal, the Likert measure used was: not disturbing (-2), slightly not disturbing (-1), neutral (0), slightly disturbing (+1), very disturbing (+2).

After recording the participant’s answer, a personality assessment test (objective test) was administered. This enabled us to collect data on the participant’s impressions of the narrator: whether they had connected with him, trusted him etc.

**Task**

For first part of the experiment, the participants were told to view a photograph, reflect upon it, and make-up a story explaining the content of the photograph. The participants were instructed to narrate the story as naturally as possible. By giving them this instruction the participants were made free to use rich gestures. Also, they were informed of the sideline videotaping activity. This part was earlier termed as the PDT test.

To help the P1 participants to come up with a story, the following open ended questions were asked: (i) what has led up to the event shown, (ii) what is happening at the moment, (iii) what the characters are feeling and thinking, and (iv) what is then the outcome of the story. The participants were told to author their own story but were explicitly instructed not to describe or refer to the PDT test stimuli picture in detail. This was to ensure that the narration and the method of narration contained clues about the picture which could only be unlocked by carefully listening to the narrative.

Furthermore, for P1, the participants were told to be free, informal, alive, animated and real. And, they were asked to pretend that they were narrating their story to a friend in English.

For P2, the participants were instructed to either view or hear a video that was being played for them. Post the video viewing, from a collage of photos, the P2 participants were instructed to select only one photo which best fits the story. After recording the answer, the participants were asked to answer a small questionnaire which gauged the quality of media shown.

This process was to be repeated six times to cover all the six narratives. After the final sixth narrative, the participants were asked to do a personality assessment test on the instructor. We had used the HEXACO personality inventory instrument for this purpose [10].

**RESULTS**

To correlate the results to the hypothesis, the following additional data points were monitored: (1) the number of accurate selections of the photo stimuli used for the PDT test in P1 which is an indication of proper comprehension, or better yet, an indication that there was no functional message loss in transmission; (2) values assessing the media and the channel quality this is an indication of the interference of the transmission format with the message (as noticed and felt by the message receivers); (3) values that assessed the P1 narrator at a social level this is an indication of how the loss of non-verbal information like gestures has impacted the message content or delivery, or even the messenger (i.e. the narrator) itself.

Results show that overall 82 respondents were able to pick the correct PDT test photo from a collage featuring multiple decoys (Table 2). That is, over 46% of the overall sample population identified the photo successfully. The group that got the untreated raw video had 23 (38%) of 60 accurate selections. The group that got the cropped video got 30 (50%) of the 60 accurate selections. And, the group that got only the audio of P1 got 29 (48%) of 60 identifications.

The participants on average rated the quality of the processed video to be at +0.4, which fell in-between the Likert ’neu-
Table 2. Results of Message Comprehension Test in P2.

<table>
<thead>
<tr>
<th></th>
<th>Full raw video</th>
<th>Video cropped to passport size</th>
<th>Audio &amp; no video</th>
<th>Overall</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of accurate photo identifications</td>
<td>23</td>
<td>30</td>
<td>29</td>
<td>82</td>
<td>3.1</td>
</tr>
<tr>
<td>% of accurate photo identifications</td>
<td>38%</td>
<td>50%</td>
<td>48%</td>
<td>46%</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 3. Results of evaluating the quality of media used for the 6 narratives in P2.

Table 4. Results of Hexaco scores from P2.

DISCUSSION

From the results it is evident that in our experiment, less is indeed more. Apparently, participants with cropped video, and those with just audio, were not only getting the message being communicated, but were doing better than those who got to see the full, unedited, raw video. In fact, if one of the outliers is removed (one corresponding to PDT photo stimuli 3), amongst all the 6 narratives in the P1, we find that the average accurate photo identification rate jumps up by 8% and makes the cropped and just-audio groups’ photo identification figures reach a score of 58%. This indicates two things to us: (1) the message content does play a role in getting the gist of the message across, and (2) showing a mid-level shot of the narrator with all the background and body plus hand gesture need not add significant additional value to the message content.

When it came to evaluating the narrator himself, we found that the HEXACO ratings were no less favorable. That is, receivers of the message did not feel like they were any less comfortable, or any less bonded with the narrator in all three treatments. In the dimensions of Sincerity, Social Boldness, Sociability, Liveliness, Organization, Perfectionism, Aesthetic Appreciation, Inquisitiveness, Creativity, Unconventionality, there was not much deviation implying that, in rating the narrator or in bonding with the virtual presence of the messenger, it did not matter which format they received story in.

Given that cropping of pre-recorded video lectures is still communicating the same message and given that, even without the non-verbals the videos are still maintaining the social connect between the instructor and the students, we propose cropping out the non-content relevant background visuals, hand gestures, body gestures etc. With this, for transmission, we calculate a raw bit rate savings of 7,55,100 bits per frame, over 24 frames per second, resulting in 972.2 Giga bits savings in a one hour lecture. Of course, with proper codec and good transmission protocol, this bit rate is tremendously reduced in addition to this saving we have demonstrated through our perceptually driven analysis.
CONCLUSION
From the experimental results we can conclude that the transmission savings are achievable. Moreover, this does not come at a cost of compromising on the message being transmitted or loosing on the emotional or social component of the professor who is authoring the content. In addition, cropping for educational videos seems to also increase the focus the student is giving to the message of the video. In the case of course-content videos, many videos can now be post-processed by this demonstrated cropping method. If the videos are not entirely talking head types, then only those portions which are having talking head imagery can be identified and cropped. This too will reduce video’s footprint and transmission cost.

Finally, we conclude that the earlier stated socio-cultural constraints of not having good infrastructure in tier 2 and tier 3 educational institutes and also students wanting to watch videos on their mobile phones, are practical requirements that need to be considered in design, development and deployment of NPTEL like digital assets. This documentation of our research effort, remains as just one specific study of a potential non-technical solution that can be explored in helping users interact better with large size course-content educational videos.

REFERENCES
15. Ministry of HRD, G. o. i. National program on technology enhanced learning @ONLINE, Feb. 2012.