Utilising High-Level Features in Summarisation of Academic Presentations

Keith Curtis
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
Keith.Curtis@AdaptCentre.ie

Gareth J.F. Jones
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
Gareth.Jones@AdaptCentre.ie

Nick Campbell
ADAPT Centre
School of Computer
Science & Statistics
Trinity College Dublin
Dublin, Ireland
nick@tcd.ie

ABSTRACT

We present a novel method for the generation of automatic video summaries of academic presentations. We base our investigation on a corpus of multimodal academic conference presentations combining transcripts with paralinguistic multimodal features. We first generate summaries based on keywords by using transcripts created using automatic speech recognition (ASR). Start and end times for each spoken phrase are identified from the ASR transcript, then a value for each phrase created. Spoken phrases are then augmented by incorporating scores for human annotation of paralinguistic features. These features measure audience engagement, comprehension and speaker emphasis. We evaluate the effectiveness of summaries generated for individual presentations, created using speech transcripts and paralinguistic multimodal features, by performing eye-tracking evaluation of participants as they watch summaries and full presentations, and by questionnaire of participants upon completion of eye-tracking studies. Summaries were also evaluated for effectiveness by performing comparisons with an enhanced digital video browser.

CCS CONCEPTS

•**Computing methodologies** →**Video summarization;** Classification and regression trees; •**General and reference** →*Evaluation;*

KEYWORDS

Video summarisation; classification; evaluation; eye-tracking

ACM Reference format:
Keith Curtis, Gareth J.F. Jones, and Nick Campbell. 2017. Utilising High-Level Features in Summarisation
of Academic Presentations. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 7 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079028

1 INTRODUCTION

Archives of multimedia content are currently growing very rapidly. Every minute in 2016, 300 hours of new video material was uploaded to YouTube, while 2.78 million videos were viewed, skype users made 104,300 calls, 38,050 hours of music were listened to on spotify and 38,194 new posts were made to Instagram. It is a growing challenge for users to be able to locate and browse content of interest in such multimedia archives - either in response to user queries or in more free or informal exploration of content. Browsing multimedia archives to find relevant information is extremely time consuming. While this is a challenging problem for multimedia including a video channel, this is particularly challenging for spoken content where a user must listen to the spoken audio track in a linear mode. This represents a growing challenge for accessing content considered relevant or of interest to users of multimedia archives where the significant information is in the audio of the multimedia file such as lectures, presentations etc. of which archives are growing rapidly.

The goal of this work is to provide an effective and efficient way to summarise audio-visual recordings where the significant information is in the audio stream. Existing multimedia information retrieval research has focused primarily on matching text queries against written meta-data or transcribed audio [2], in addition to seeking to match visual queries to low-level features such as colours and textures and shape and object recognition and person recognition, plus scene type classification - urban, countryside, and named individuals or places etc. [6]. In the case of multimedia content where the information is primarily visual, the content can be represented by multiple key-frames extracted using methods such as object recognition and facial detection. These are then matched against visual queries, and retrieved videos shown to the user using key-frame surrogates, with the user playing back in full videos which they believe to be relevant to the information need. Matching of the visual component of these queries is generally complemented by textual search against a transcript of any available spoken audio and any meta-data provided [9].

In this work we develop a novel video presentation summarisation algorithm to automatically generate summaries of academic presentations using classified areas of high audience engagement, audience comprehension, and areas of intentional or unintentional emphasis applied by the speaker. We address the following research question 'Can areas of special emphasis provided by the speaker, combined with detected areas of high audience engagement and

ICMR '17, June 6–9, 2017, Bucharest, Romania
© 2017 ACM. ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079028

high levels of audience comprehension, be used for effective summarisation of presentations?' We evaluate this summarisation approach using eye-tracking and by questionnaire. Our summarisation approach is also evaluated by comparisons to an enhanced digital video browser for quickly skimming presentations.

In section 2 of this paper we introduce relevant related work in video summarisation. Section 3 introduces the multimodal corpus used for these experiments while section 4 describes the classification of high level features used for creating these summaries. Section 5 describes the procedure for creating automatic summaries followed by section 6 which describes the evaluation tasks performed and results of these. Conclusions are offered in section 7 of the paper.

2 PREVIOUS WORK

This section looks at related work for the summarisation and skimming of academic presentations.

Ju et al. [8] present a system for analysing and annotating video sequences of technical talks. They use a robust motion estimation technique to detect key frames and segment the video into subsequences containing a single slide. Sub-sequences are stabilised to remove motion. Potential gestures are tracked using active contours. By successfully recognising all pointing gestures the authors are able to fully annotate presentations per slide. This first automatic video analysis system helps users to access presentation videos intelligently.

He et al. [5] use prosodic information from the audio stream to identify speaker emphasis during presentations, in addition to pause information to avoid selecting segments which start mid-phrase. They also garner information from slide transition points to indicate the introduction of a new topic or sub-topic. They develop three summarisation algorithms for slide transition based summary, pitch activity based summary and summary based on slide, pitch and user-access information. They use surveys for evaluation and find that computer generated summaries were rated poorly on coherence 5.3 to 3.5 in which participants complained that summaries jumped topics. In this work they also found that audio-visual presentations were less susceptible than the audio-only stream to pitch based emphasis analysis. This is the first work attempting to summarise presentation video by speaker emphasis. In this they found that speaker emphasis was not enough information to generate effective summaries.

Li et al. [10] develop and evaluate an enhanced digital video browser for browsing through different categories of digital video. The categories evaluated are Classroom, Conference, Sports, Shows, News and Travel. They evaluate the effectiveness of the following enhanced browser controls - Time Compression (TC), Pause Removal (PR), Table of Contents (TOC), Shot Boundary (SB), Timeline markers and jump controls. For browsing of conference presentations, TC and PR were found to be the most effective tools for improved video browsing with scores of 6.9 and 6.5 out of 7 respectively. This enhanced digital video browser gives us additional comparison method to evaluate the effectiveness of generated presentation summaries. From this work we gain the idea of an additional evaluation measure by comparison with the enhanced digital video browser.

Joho et al. [7] capture and analyse the user's facial expressions for the generation of perception-based summaries which exploit the viewer's affective state, perceived excitement and attention. Perception-based approaches are designed to overcome the semantic gap problem in summarisation by finding affective scenes in video. They find it unlikely that a single summary could be commonly seen as the highlight of videos by viewers. Results suggest that there were two or three at least distinguished parts of videos that can be seen as the highlight by various viewers.

Work by Pavel et al. [11] created a set of tools for creating video digests of informational video. Informal evaluation suggests these tools make it easier for authors of informational talks to create video digests. They also found that video digests afford browsing and skimming better than alternative video presentation techniques.

3 MULTIMODAL CORPUS

No standard publicly dataset exists for work of this nature, since we require recordings of the audience and of the speaker for academic presentations. For this work we used the International Speech Conference Multi-modal corpus (ISCM) from Curtis et al. [3]. This contains 31 academic presentations totalling 520 minutes of video, with high quality 1080p parallel video recordings of both the speaker and the audience to each presentation. We chose four videos from this dataset for our evaluation of our video summaries.

4 CLASSIFICATION OF HIGH-LEVEL FEATURES FOR SUMMARISATION

In this work, we develop a novel video summarisation algorithm to generate video summaries from the classification of high-level subjective concepts such as audience engagement, comprehension and speaker emphasis. In earlier work we developed effective technologies for automatic classification of these features. The following subsections briefly review this work and summarise our classification results on the ISCM corpus.

4.1 Classification of Audience Engagement

Prediction of audience engagement levels and applying ratings for 'good' speaking techniques [3] were performed by first employing human annotators to watch video segments of conference presentations and to provide a rating on an ordinal scale of how good that speaker was at presenting the material. Audience engagement levels were measured in a similar manner by having annotators watch video segments of the audience to academic presentations and to provide an estimate of just how engaged the audience appeared to be as a whole. Classifiers were trained on extracted audio-visual features using an Ordinal Class Classifier.

The qualities of a 'good' speaker can be predicted to an accuracy of 73% over a 4-class scale. Using speaker-based techniques alone, audience engagement levels can be predicted to an accuracy of 68% over the same scale. By combining with basic visual features from the audience as whole, this can be improved to 70% accuracy.

4.2 Identification of Emphasized Speech

Identification of intentional or unintentional emphasized speech [1] was performed by having human annotators label areas of emphasised speech. Human annotators were asked to watch through short video clips and to mark areas where they considered the speech to be emphasized, either intentionally or unintentionally. Analysis was performed on this data to attempt to identify existing conditions for areas in which annotators had marked emphasis. Basic audio-visual features of audio pitch and visual motion were then extracted from this data. From the analysis performed, it was clear that speaker emphasis occurred during areas of high pitch, but also during areas of high visual motion coinciding with areas of high pitch.

Candidate emphasised regions were marked from extracted areas of pitch within the top 1, 5, and top 20 percentile of pitch values, in addition to top 20 percentile of gesticulation down to the top 40 percentile of values respectively. All annotated areas of emphasis contained significant gesturing in addition to pitch with the top 20 percentile.

4.3 Predicting the Speakers Potential to be Comprehended

Prediction of audience comprehension [4] was performed. Human annotators were recruited through the use of crowdsourcing. Annotators were asked to watch each section of a presentation and to first provide a textual summary of the contents of that section of the presentation and following this to provide an estimate of how much they comprehended the material during that section. Audio-visual features were extracted from video of the presenter in addition to visual features extracted from video of the audience, and OCR over the slides for each presentation. Additional fluency features were also extracted from the speaker audio. Using the above described extracted features, a classifier was trained to predict the speakers potential to be comprehended.

We showed that it is possible to build a classifier to predict potential audience comprehension levels, obtaining accuracy over a 7-class range of 52.9%, and over a binary classification problem to 85.4%.

For the purpose of investigating the effect of these features on summarisation, in this study we rely on manual labels of these features within the investigated presentation recordings. This enables us to concentrate on the impact of the features without needing to consider the impact of errors in automatic feature detection.

5 CREATION OF PRESENTATION SUMMARIES

This section describes the steps involved in the generation of presentation summaries. Summaries are generated using transcripts created using automatic speech recognition (ASR), keywords, and annotated values for 'good' public speaking techniques, audience engagement, intentional or unintentional speaker emphasis and the speakers potential to be comprehended. Using the ASR outputs, we use pause information derived from the ASR output, which gives start and end times for each spoken phrase during the presentation. This provides a basis for the separation of presentations at the phrase-level. We first apply a ranking for each phrase based on the number of keywords, or words of significance, contained within it. For the first set of baseline summaries, we generate summaries by using the highest ranking phrases.

Additional ranking is applied to each sentence based on human annotations of 'good' speaking techniques. Combination of our baseline scores for each phrase with other features is carried out using a weighted linear summation of the available features. In the absence of a suitable training set to set optimal values of the summation weights, we select the weights based on our observed intuitions of their behaviour and significance Speaker Rating's are halved before applying this ranking to each phrase. We half the values for speaker ratings so as not to overvalue this feature, as these values are already encompassed for classification of audience engagement levels. Following this, audience engagement annotations are also applied directly. We take the true annotated engagement level and apply this ranking to each sentence contained within each segment throughout the presentation. As emphasis was not annotated for all videos, we use automatic classifications for intentional or unintentional speaker emphasis. For each classification of emphasis, we apply an additional ranking of 1 to the phrase containing that emphasised part of speech. Finally, we use the human annotated values for audience comprehension throughout the data-set. Once again the final comprehension value for each segment is also applied to each sentence within that segment. For weightings, we choose to half the Speaker Rating annotation, while choosing to keep the original for other annotations, this is to reduce the impact of Speaker Ratings on the final output. Points of emphasis receive a ranking of 1, while keywords receive a ranking of two, in order to prioritise the role of keywords in the summary generation process.

To generate the final set of video summaries, the highest ranking phrases in the set are selected. To achieve this, the final ranking for each phrase is normalised to between 0 and 1. By then assigned an initial threshold value of 0.9, and reducing this by 0.3 on each iteration, we select each sentence with a ranking above that threshold value. By calculating the length of each selected sentence, we can apply a minimum size to our generated video summaries. Final selected segments are then joined together to generate small, medium and large summaries for each presentation.

Algorithm 1 Generate Summaries

```
for all $Sentence \rightarrow S$ do
    if $S\_contains\_Keyword$ then
        $S \leftarrow S + 2$
    end if
    $Engagement \rightarrow E$
    $SpeakerRating \rightarrow SR$
    $Emphasis \rightarrow Es$
    $Comprehension \rightarrow C$
    $S \leftarrow S + E$
    $S \leftarrow S + SR/2$
    $S \leftarrow S + Es$
    $S \leftarrow S + C$
end for
while $Summary < length$ do
    if $S \geq Threshold$ then
        $Summary \leftarrow S$
    end if
end while
```

6 EVALUATION OF VIDEO SUMMARIES

In this study we hypothesise that the classification of areas of 'good' speaking techniques, audience engagement, intentional or unintentional speaker emphasis, and the speakers potential to be comprehended by the audience, can be used to improve summarisation methods for academic presentations. To evaluate this, we first crowdsource human evaluation tasks by asking workers to watch summaries generated using keywords alone and summaries generated using all available information. Additional evaluation of automatically generated presentation summaries is performed by eye-tracking of participants as they watch a full presentation video and an automatic video summary. The question being addressed by eye-tracking of participants, is whether or not participants spend a larger proportion of the time viewing slides during summaries than full presentations. A complimentary comparison is also performed against the use of an enhanced digital video browser, which removed pauses and allowed playback to be increased to 250% of the norm to support manual skimming, these were previously found to be very effective for gaining a quick overview of conference presentations.

6.1 Evaluation Between Summary Types

Table 1 lists results of the comparison between summaries built using keyword information only and summaries built using all available information - keywords in addition to classifier outputs of the concepts of engagement, 'good' speaking techniques, comprehension and emphasis. A total of 48 participants watched the summaries and answered the questionnaire on each summary. The questions addressed in this comparison are listed in the following section below, and answers are ranked on a Likert scale from 1 to 7.

Table 1: Total ranking per summary type for each video

Video	Q1	Q2	Q3	Q4	Q5
plen2_Key	3.5	5	4	5	4
	42	65	44	55	44
plen2_All	3.5	5	4	4	4
	42	61	43	52	43
prp1_Key	3.42	5	3	4	3
	41	58	38	46	38
prp1_All	3.42	5	3	4	3
	41	56	39	46	39
prp5_Key	5.08	6	5	5	5
	61	67	57	56	57
prp5_All	4.25	6	5	5	5
	51	66	58	55	58
spRT6_Key	3.5	5	3	3	3
	42	55	39	39	39
spRT6_All	3.42	5	3	4	3
	41	59	31	49	31

These results indicate that user's perceive little difference overall, between summaries generated using keyword information only, as summaries generated using all available information. The most notable results indicate that users perceive the keyword summary of prp_5 to be easier to understand than the summary created using all available information for this video. Interestingly, user's also perceive the summary generated using all available information from speechRT6 video to be much more coherent than the summary generated using only keyword information from this video. This may however be explained by the fact that speechRT6 was discovered by our classifiers to be less engaging and comprehensible than other videos in the collection, while prp_5 was found to be the most engaging video. This would indicate that the effectiveness of using engagement information depends on just how engaging the video was in the first place.

6.2 Gaze-Detection Evaluation

Eye-tracking was performed for evaluation of presentation summaries. Participants watched one full conference presentation whilst having their eye-movements tracked and answered a questionnaire on the same. Participants also watched a separate presentation summary, again whilst having their eye-movements tracked and answered a number of questions on same. The question being addressed here was whether or not participants spent a larger proportion of the time viewing the slides during summaries than full presentations, to support the hypothesis that summaries contain a higher concentration of new information than full videos.

Table 2: Eye-tracking video 1 slides scene by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.02	0.05	0.655
summ	**full**	**percent**	**Scheffe**	**11.07**	**4.97**	**0.043**
summ	**full**	**FCp100**	**Scheffe**	**37.32**	**14.32**	**0.021**

Table 2 shows a statistically significant difference between the summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. As also demonstrated in the core figures in table 5, this indicates a significant difference in the amount of time users spent fixating on the slides while watching the summary than the full video. This indicates that users found there to be a much higher concentration of new information during the summary than during the slides.

Table 3: Eye-tracking video 2 slides scene by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.08	0.08	0.337
summ	full	percent	Scheffe	-2.72	7.90	0.735
summ	full	FCp100	Scheffe	7.04	11.16	0.538

Again in Table 3, key differences are observed in the average fixation duration per scene, and to a lesser extent in the fixation count per 100 seconds, more clearly visible from the figures in Table 5, neither of these differences are statistically significant.

Table 4 shows that there is a significant difference in the mean fixation length over the slides during the summary than during the full presentation. As can be seen by looking at the core figures in Table 5, the averaged fixation duration is much shorter for fixations on slides. This indicates that users tend to read over slides in order to follow and absorb printed information.

Table 4: Eye-tracking video 3 slides scene by version

I	J	Variable	Measure	Diff	Error	Sig
summ	**full**	**FD.M**	**Scheffe**	**-0.15**	**0.05**	**0.009**
summ	full	percent	Scheffe	-2.67	6.54	0.689
summ	full	FCper100	Scheffe	16.91	13.54	0.232

Table 5: Eye-tracking video 4 slides scene by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.06	0.05	0.266
summ	**full**	**percent**	**Scheffe**	**15.68**	**6.42**	**0.028**
summ	**full**	**FCp100**	**Scheffe**	**54.96**	**16.98**	**0.006**

Table 5 shows a statistically significant difference between the summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. As demonstrated in the core figures in Table 5, this indicates a significant difference in the amount of time users spent fixating on the slides while watching the summary than the full video. This indicates that users found there to be a much higher concentration of new information during the summary than the full version.

Table 6: Totals per vid, version, scene (V.V.S = Vid,Version,Scene)

V.V.S.	FD.N	DF.M	FD.S	%	FC.S	FCp100
1.1.1	354.875	0.4213	143.874	66.608	354.875	164.294
1.1.2	62.5	0.9588	55.296	25.600	62.5	28.935
1.2.1	1223.5	0.4463	536.485	55.537	1223.5	126.978
1.2.2	299.5	1.2463	329.468	34.106	299.5	31.004
2.1.1	248.125	0.59	145.479	66.127	248.125	112.784
2.1.2	45.125	1.2438	6.518	25.69	45.125	20.511
2.2.1	923.125	0.6663	601.053	68.849	923.125	105.742
2.2.2	165.5	1.0775	160.774	18.416	165.5	18.958
3.1.1	140.125	0.4288	60.123	40.623	140.125	94.679
3.1.2	53.75	1.3375	69.969	47.276	53.75	36.318
3.2.1	596.5	0.58	332.088	43.297	596.5	77.771
3.2.2	294.875	1.3288	324.413	42.296	294.875	38.445
4.1.1	345.125	0.4475	149.583	79.144	345.125	182.606
4.1.2	40.375	0.6113	24.523	12.975	40.375	21.362
4.2.1	1194.75	0.505	594.019	63.464	1194.75	127.644
4.2.2	219.625	0.8137	181.351	19.375	219.625	23.464

Table 6 shows the averaged core values for eye-tracking measurements per video, version and scene, from which the above tables 2 - 5 are calculated. For this, version denotes either a summarisation video or the full video, whereas scene denotes either the area of the slides or the area around the speaker themselves. Measurements obtained include number of Fixations, Mean length of fixations, Total Sum of fixation lengths, percentage of time fixated per scene, average number of fixation counts and Number of fixation counts per 100 seconds.

6.3 Questionnaire for Eye-Tracking Participants

Following the participant's completion of eye-tracking experiment, they completed a questionnaire. This questionnaire asked participants how much they agreed with each of the 5 statements issued, depending on whether they watched a summary or the full video, on a 5-Likert scale.

The following is the five statements on each video for which participants indicated their level of agreement:

(1) The video / summary is easy to understand.
(2) The video / summary is enjoyable.
(3) The video / summary is informative.
(4) The video / summary is coherent.
(5) I would have preferred to see a summary of this video. / This would aid in deciding whether to watch the full video.

Participants indicated their level of agreement with each of the five statements on a 5-point Likert scale, from Strongly Disagree to Strongly Agree.

The following is the average rankings for each video based on answers given to the questionnaire listed above. Video summaries were rated as being slightly less easy to understand than full presentation videos, with the largest different being for video 2, prp1. All video summaries have also received high ratings for their ability to help users to decide whether they wished to watch the full presentation video. Summarisation videos all received good ranking for informativeness, with all videos receiving a ranking of over 2.5, the best of these is again video 2, prp 1, which received an overall informative ranking of 3.22. The video found to be the least informative was video 4, speechRT6. This video was chosen for summarisation evaluation as it had been found by our classifiers to be the least engaging and least comprehensible presentation from the collection. This video summary still received an overall informativeness ranking of 2.5 out of 5, indicating that this video summarisation strategy helps maintain coherence of presentation videos when generating summaries.

Most summaries evaluated were found to be slightly less enjoyable than their original. The exception to this is video 4, speechRT6. As explained previously this video was found by our classifiers to be the least engaging and least comprehensible video from the collection. That this video summary is found to be more enjoyable than in its original form indicates that watching summaries can be a good alternative to watching full presentations of which the user may not be very interested.

A word of warning with these results is that none of these results indicated below have shown to be statistically significant, limiting the conclusions which can be drawn from these results.

6.4 Evaluation by Comparison with Enhanced Digital Video Browser

For final evaluation of presentation summaries, we performed a comparison against the properties of an enhanced digital video browser as studied by [10]. In this work, they evaluated the effectiveness and most useful features of an enhanced digital video browser. Within the domain of conference presentations they found

Table 7: Averaged rankings per video

Video	Q1-avg	Q2-avg	Q3-avg	Q4-avg	Q5-avg
vid-1 (s)	1.75	1.75	2.625	2.625	3.75
vid-1 (f)	2.5	2.75	4	3.5	3.75
vid-2 (s)	2.8889	2.8889	3.2222	2.8889	4.3333
vid-2 (f)	4.5	4	3.7	4.1	4.1
vid-3 (s)	3	3.25	3.125	3.5	4.125
vid-3 (f)	4.125	3.625	4.5	4.5	3.125
vid-4 (s)	1.7	1.8	2.5	2.4	3.8
vid-4 (f)	2.3333	1.6667	3.3333	3.3333	4.2222

Table 8: Averaged rankings per summary or full video

Video	Q1-avg	Q2-avg	Q3-avg	Q4-avg	Q5-avg
Summ	2.3143	2.4	2.8571	2.8286	4
Full	3.4	3.0286	3.8571	3.8571	3.8286

that Time Compression and Pause Removal were rated by participants as the most useful features. For further evaluation of automatically generated presentation summaries, we compare against the use of an enhanced digital video browser providing the features of Pause Removal and Time Compression.

For these evaluations, participants were given 5 minutes to gain a clear and concise overview of a presentation which they had previously missed, in order to be able to take part in a meeting discussing what was presented. Participants were given an enhanced digital video browser which removed pauses and allowed participants to increase or decrease playback rate as required, up to a maximum of 250% normal speed. Using this, participants had 5 minutes to gain a clear and concise overview of one of the presentations, and then answered 3 questions on the use of this tool. Participants were also given an automatically generated summary of the presentation. This summary was under 5 minutes in length and participants were again given 3 questions to answer on the use of this tool.

The following is the three statements on each tool for which participants indicated their level of agreement:

(1) This tool is easy to use.
(2) This tool allowed me to gain a clear and concise overview of the presentation.
(3) This would be my tool of choice for tasks of this nature.

Table 9: Averaged rankings per video

Video	Q1-avg	Q2-avg	Q3-avg
video-1 summ	4.5625	4.375	3.625
video-1 enh	4.5	3.9375	3.9375
video-2 summ	4.75	4.6875	4.1875
video-2 enh	5.125	5.1875	4.75
video-3 summ	4.875	4.6875	4.75
video-3 enh	4.125	4.3125	4.375
video-4 summ	4.8125	4	4
video-4 enh	4.4375	4.0625	3.9375

Results in table 9 indicate that participants were found to be 4.5% more likely to find that generated summaries were easier to use than the enhanced digital video browser for gaining a quick, clear and concise overview of missed presentations. Participants were also 1.5% more likely to agree that automatically generated video summaries were better for gaining a clear and concise overview of missed presentations than was the use of enhanced digital video browser. However, they were still 2.6% more likely to choose the enhanced digital video browser as their tool of choice.

These results are not however statistically significant, meaning that effectively we ended up with very equal rankings between automatically generated video summaries and and the use of enhanced digital video browser for gaining a quick, clear and concise overview of missed presentations.

7 CONCLUSIONS

The results of eye-tracking experiments performed in this study indicate that generated summaries tend to contain a higher concentration of relevant information than full presentations, as indicated by the higher proportion of time participants spend carefully reading slides during summaries than during full presentations, and also by the lower proportion of time spent fixating on the speaker during summaries than during full presentations. We also find that watching generated presentation summaries is a good alternative to watching full presentations in which the viewer might not be fully interested. However the questionnaire did find that generated video summaries receive lower ratings than full video presentation for being easy to understand, enjoyable, informative and coherent.

More encouraging results are shown by our comparisons with an enhanced digital video browser, which show that our generated summaries compare quite favourably to the use of an enhanced digital video browser, though these results lack statistical significance. In this case generated video summaries receive slightly higher scores than an enhanced digital video browser for 'ease of use' and effectiveness for 'gaining a clear and concise overview of the presentation'. However, they receive slightly lower scores for 'tool of choice'. Given that the work of [10] showed that an enhanced digital video browser scores very favourably for skimming presentation videos, and our own comparisons which rate our generated summaries at least equal in terms of effectiveness, this is a very promising result for these summarisation tasks.

The results of this study indicate that while classification of areas of engagement, emphasis and comprehension can be useful for summarisation, this may depend on how engaging and comprehensible videos were in the first place.

While this work demonstrates potential to exploit non-verbal features effectively in summarisation of presentation videos, we need to conduct further work to more fully understand the potential contributions of each of these features in individual videos, and their potential utility in comparison to the use of transcripts and other potentially text annotations of the content.

ACKNOWLEDGMENTS

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University.

REFERENCES

[1] J. Allwood, E. Ahlsén, S. Lanzini, A. Attaran, K. Curtis, G. J. Jones, N. Campbell, J. Frid, G. Ambrazaitis, M. S. Lundmark, et al. The 4th european and 7th nordic symposium on multimodal communication (mmsym 2016), copenhagen 29-30 september 2016: Extended abstracts. pages 6–7, 2016.

[2] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon. Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 105–112. ACM, 2008.

[3] K. Curtis, G. J. Jones, and N. Campbell. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 35–42. ACM, 2015.

[4] K. Curtis, G. J. Jones, and N. Campbell. Speaker impact on audience comprehension for academic presentations. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 129–136. ACM, 2016.

[5] L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 489–498. ACM, 1999.

[6] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, pages 527–536. ACM, 2010.

[7] H. Joho, J. M. Jose, R. Valenti, and N. Sebe. Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 31. ACM, 2009.

[8] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber. Summarization of videotaped presentations: automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):686–696, 1998.

[9] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19, 2006.

[10] F. C. Li, A. Gupta, E. Sanocki, L.-w. He, and Y. Rui. Browsing digital video. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 169–176. ACM, 2000.

[11] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala. Video digests: a browsable, skimmable format for informational lecture videos. In *UIST*, pages 573–582, 2014.