Corpus Generation and Analysis: Incorporating Audio Data Towards Curbing Missing Information

Atiqah Izzati Masrani and Yoshihiko Gotoh

University of Sheffield, United Kingdom
{amasrani1,y.gotoh}@sheffield.ac.uk

Abstract. As video data becomes widely available, it is crucial that these videos are properly annotated for effective search, mining and retrieval purposes. Significant work has been done to explore natural language description as it can provide better understanding of the video content. Ideally, a summary should be informative and accurate in order for the users to have good understanding of the video content. An experiment has been conducted to evaluate the impact of audio information towards natural language summary annotations of a video content. The experiment proved that although events and human activities can be captured using visual features alone, key information of the video content would be missing without the audio information. Thus, future work on natural language summary generation should incorporate both visual and audio data to curb missing and erroneous information.

Keywords: Corpus generation, hand annotation, visual features, audio data

1 Introduction

Nowadays, there is an abundant of videos that are accessible online. The widespread use of the Internet has allowed videos to be accessed easily via video search engines such as from the YouTube or Daily Motion. The YouTube itself has more than 1 billion users and is estimated to have 300 hours of video uploaded every minute. It generates billions of views on a daily basis. Furthermore, the number of hours people spent on watching YouTube each month is up 50% year over year. [1] This raises the question on how the users can be more selective during video browsing and retrieval. Although some of these videos are well-organized with manually annotated tags or labels, some has no clear description of its content. Therefore, users may tend to skim through the video to grasp a hint of its semantic content.

Video summarization addresses this issue by providing brief information of the video. Significant work has been done in this area with a large part of it optimizes graphical representations. The graphical representations can be further

[1] https://www.youtube.com/yt/press/statistics.html

divided into two classes. The first class focuses on compressing the video into a shorter representation of the video that is also known as video skimming. This include works from [1], [2] and [3]. The second class uses image key frames extracted from the video stream to reflect the content or the highlights of the video [4] [5].

Natural language has also proven to be a popular choice to represent a video. It is an appealing option as it is less space consuming, has faster processing time for retrieval and is readable by both human and machine. Most early research such as works by [6] and [7] uses representative keywords. Using keywords can boost the potential for fast video retrieval because it helps efficient video categorization. However, using keywords alone may not be able to capture the whole key points of the video, as keywords tend to be ambiguous. This may affect the accuracy and effectiveness of video classification due to its ambiguity and lack of information. Natural language representation in the form of a "summary" or "abstract" is one way to address this. There has been significant works on creating a natural language summary that emphasizes on its coherency and informativeness. However, human's perspective when watching a video is subjective. Although presented with the same visual scene, one's interpretation may vary. This may influence on how they will write the summary of the video.

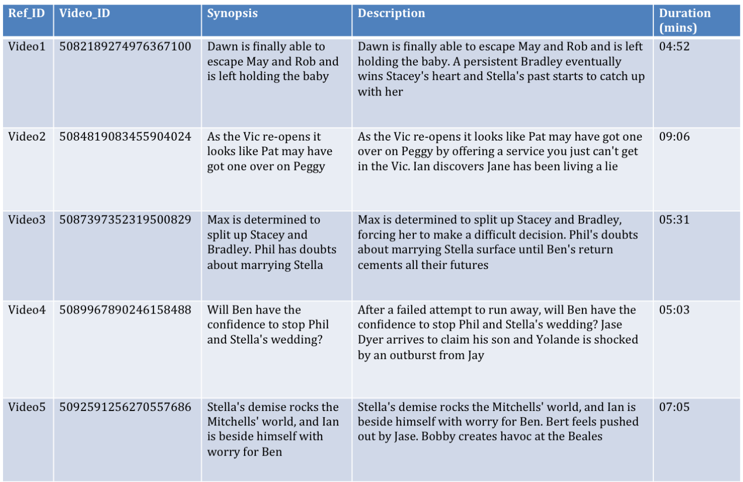
In this paper, an experiment was conducted to study the overlapped similarities of human's perspectives and also the impact of incorporating audio data during summary annotations. This paper aims to prove that the dissimilarity lies in the words used to semantically convey the meaning, and the similarity lies in the key information that is included in the summary. This paper also aims at proving that both the visual and audio data are important towards determining the key points of the video. Thus, using only one without the other towards natural language generation framework for video data may cause missing or erroneous information in the summary.

2 Corpus Generation

As video data becomes widely available, it is crucial that these videos are properly annotated for effective search, mining and retrieval purposes. Significant progress has been made to use natural language description as it can provide better understanding of the video content such as work by [9], [10], and [11]. Most of these works crafted their own video corpora that consist of the video data and its corresponding hand annotation. Each dataset are specifically designed with a certain prerequisites or constraints to fulfill a specific task or purpose.

In [8], the dataset is designed for the task of generating natural language descriptions of the video content. The work focuses on the natural language generation phase that is heavily dependent on the visual features extracted during the HLFs processing phase. The dataset is crafted from videos that consist of subjects, objects, actions and scene settings that can be easily identified using existing visual processing techniques. Therefore, the crafted videos are short and consist of a single shot or scene with minimal activity. Some other existing video corpora are more domain-focused such as football, traffic, surveillance, cooking videos and so on [6], [12].

In this study, video clips from the BBC EastEnder series were selected. [2] It consists of approximately 244 episodes and each are associated with its own metadata and transcripts. This dataset is chosen because of its realistic elements with human subjects showing various activities, emotions and interactions with other objects. In this experiment, 5 episodes were chosen. These episodes were crosschecked with their metadata and transcripts. Each episode has a synopsis and description included in their metadata file. Assuming that the synopsis (summary) describes the highlight of its corresponding episode, these videos were cropped focusing on the episodes' highlight. The cropped video ranges between 4 to 20 minutes of playtime. Figure 1, shows the selected video with their synopsis, description and the duration of the cropped version.



Ref_ID	Video_ID	Synopsis	Description	Duration (mins)
Video1	5082189274976367100	Dawn is finally able to escape May and Rob and is left holding the baby	Dawn is finally able to escape May and Rob and is left holding the baby. A persistent Bradley eventually wins Stacey's heart and Stella's past starts to catch up with her	04:52
Video2	5084819083455904024	As the Vic re-opens it looks like Pat may have got one over on Peggy	As the Vic re-opens it looks like Pat may have got one over on Peggy by offering a service you just can't get in the Vic. Ian discovers Jane has been living a lie	09:06
Video3	5087397352319500829	Max is determined to split up Stacey and Bradley. Phil has doubts about marrying Stella	Max is determined to split up Stacey and Bradley, forcing her to make a difficult decision. Phil's doubts about marrying Stella surface until Ben's return cements all their futures	05:31
Video4	5089967890246158488	Will Ben have the confidence to stop Phil and Stella's wedding?	After a failed attempt to run away, will Ben have the confidence to stop Phil and Stella's wedding? Jase Dyer arrives to claim his son and Yolande is shocked by an outburst from Jay	05:03
Video5	5092591256270557686	Stella's demise rocks the Mitchells' world, and Ian is beside himself with worry for Ben	Stella's demise rocks the Mitchells' world, and Ian is beside himself with worry for Ben. Bert feels pushed out by Jase. Bobby creates havoc at the Beales	07:05

Fig. 1. The selected video with their synopsis and description

3 Annotation Process

Figure 2, shows that the participants conducted two rounds of writing the summaries. In the first round, the participants were asked to watch each video without the audio and write the summary. In the second round, the participants were asked to watch each video with the audio and write the summary. No specific rules were imposed on how the summaries should be written. This is because the objective of the experiment is to obtain an unbiased (although varies) perspectives of the participants. The participants were given 2 weeks to complete the experiment. The experiment is conducted to answer these research questions:

[2] http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf

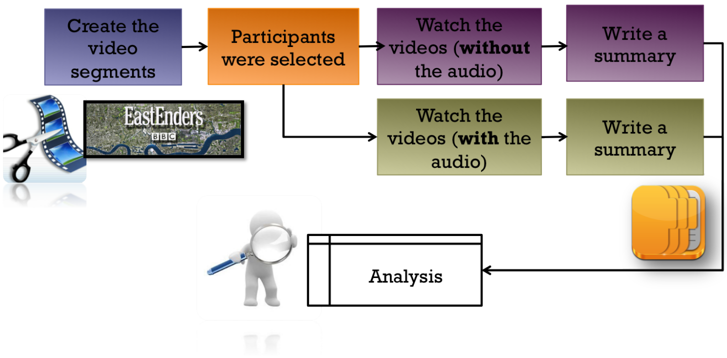

Create the video segments
Participants were selected
Watch the videos (**without** the audio)
Write a summary
Watch the videos (**with** the audio)
Write a summary
EastEnders
BBC
Analysis

Fig. 2. The experimental setup

- Can the hand annotations consist of similarities that focus on the key interest points of the video?
- Can audio data help to reduce missing information?

4 Results

Figure 3 shows a selection of image key frames that depict the highlight of Video_ID:5082189274976367100 and an example of its corresponding hand annotations from one of the participant as shown in Figure 4.

The results will be presented in two corpora that is the hand annotations without audio and the hand annotations with audio. Total number of documents for both corpora was 50 (5 participants each created 2 summaries for 5 different videos). In each corpus, two classes[3] will be manually defined:

1. Human related: gender, age, body parts, identity, emotions, grouping, dressing, actions and activities numbers.
2. Non-human related: man-made objects, natural objects, scene settings, location, colours, size

4.1 Hand Annotations (Without Audio)

Total number of documents for this corpus was 25 (5 participants each created 1 summary for 5 different videos). The total number of words for the summaries was 1856, hence the average length of one document was roughly 74 words. Total number of unique words is 402. [4]

[3] Refers to the subclasses as defined in [8]

[4] This statistic is generated using www.linguakit.com

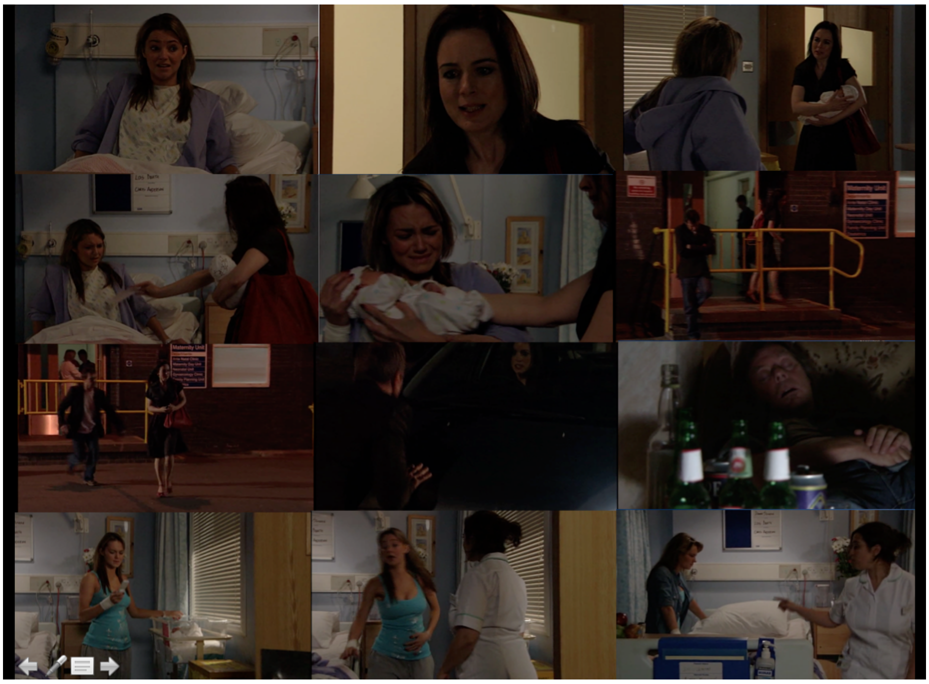
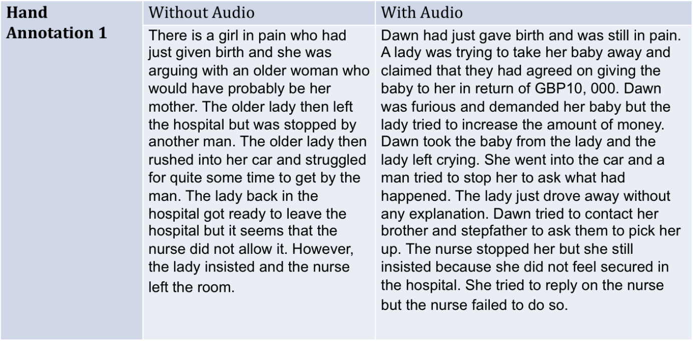


Fig. 3. A montage of the video highlights (Video_ID:5082189274976367100)

Human Related Features Figure 5 presents human related information observed in the hand annotations. The participants is shown to focus on identifying human's presence in the video because the top three most frequently used words (nouns) are woman with 41 occurrences, man with 31 occurrences and lady with 19 occurrences. For human related features, the human gender information has the highest number of occurrences: female with 77 occurrences and male with 54 occurrences. Related words such as 'lady' and 'woman' are combined into the same category 'female'. The same goes for 'male', which combines related words such as 'man' and 'boy'. Age information (e.g., old, young, child), identity (e.g., mother, nurse, groom) and grouping (e.g., one, two, crowd) are also often used. The words used to describe emotions are categorized into six basic emotions as described by Paul Ekman [5]. These six basic emotions are 'anger', 'disgust', 'fear', 'happy', 'sad', and 'surprise'. The least described features are body parts and dressing.

Non-human Related Features Figure 6 presents non-human related information observed in the hand annotations. The participants shown keen interest in identifying the location of a particular scene such as the hospital, restaurant,

[5] Paul Ekman is a psychologist and a co-discoverer of micro expressions with Friesen, Haggard and Isaacs



Hand Annotation 1	Without Audio	With Audio
	There is a girl in pain who had just given birth and she was arguing with an older woman who would have probably be her mother. The older lady then left the hospital but was stopped by another man. The older lady then rushed into her car and struggled for quite some time to get by the man. The lady back in the hospital got ready to leave the hospital but it seems that the nurse did not allow it. However, the lady insisted and the nurse left the room.	Dawn had just gave birth and was still in pain. A lady was trying to take her baby away and claimed that they had agreed on giving the baby to her in return of GBP10, 000. Dawn was furious and demanded her baby but the lady tried to increase the amount of money. Dawn took the baby from the lady and the lady left crying. She went into the car and a man tried to stop her to ask what had happened. The lady just drove away without any explanation. Dawn tried to contact her brother and stepfather to ask them to pick her up. The nurse stopped her but she still insisted because she did not feel secured in the hospital. She tried to reply on the nurse but the nurse failed to do so.

Fig. 4. An example of the hand annotation (Video_ID:5082189274976367100)

Table 1. Statistic and category frequency for the hand annotations without audio

Statistics	
Number of sentences	128
Number of words	1856
Number of unique words	402
Number of characters	8841
Characters no whitespace	7161

Category frequency	
Number of nouns	133
Number of adjectives	33
Number of verbs	129
Number of adverbs	30

church etc. They also showed interest in describing man-made objects involved (e.g., car, food, book etc.) and scene settings (e.g., ceremony, wedding, and outside). Natural objects and colours are rarely described. No word has been used to describe size.

4.2 Hand Annotations (With Audio)

Total number of documents for this corpus was 25 (5 participants each created 1 summary for 5 different videos). The total number of words for the summaries was 1983, hence the average length of one document was roughly 79 words. Total number of unique words is 426. [6]

Human Related Features Figure 7 presents human related information observed in the hand annotations. The participants is shown to focus on identifying human's presence in the video because the top three most frequently used words (nouns) are mother with 21 occurrences, baby with 20 occurrences and woman with 19 occurrences. For human related features, the human gender information has the highest number of occurrences: female with 75 occurrences and male with 75 occurrences. Identity features (e.g., mother, Dawn, nurse) also recorded high number of occurrences. Age information, emotions and grouping are described significantly. The least described features are body parts and dressing.

[6] This statistic is generated using www.linguakit.com

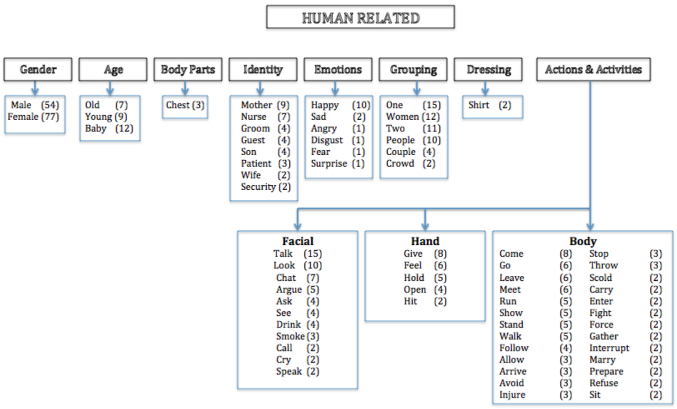


Fig. 5. Human related features in the hand annotations (without audio)

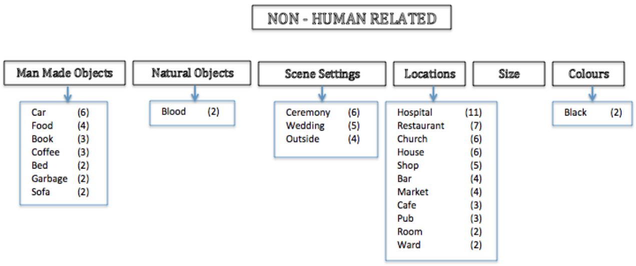


Fig. 6. Non-human related features in the hand annotations (without audio)

Non-human Related Features Figure 8 presents non-human related information observed in the hand annotations. The participants showed keen interest in identifying the location of a particular scene such as the pub, house, hospital etc. They also showed interest in describing man-made objects involved (e.g., rubbish, coffee, car etc.) and scene settings (e.g., ceremony, wedding, and outside). Natural objects and size are rarely described. No words have been used to describe colours.

5 Analysis and Discussion

The findings and analysis from this experiment will be presented in two subsections focusing on the two research questions.

Table 2. Statistics and category frequency for the hand annotations with audio

Statistics	
Number of sentences	127
Number of words	1983
Number of unique words	426
Number of characters	9500
Characters no whitespace	7713

Category frequency	
Number of nouns	138
Number of adjectives	31
Number of verbs	143
Number of adverbs	28

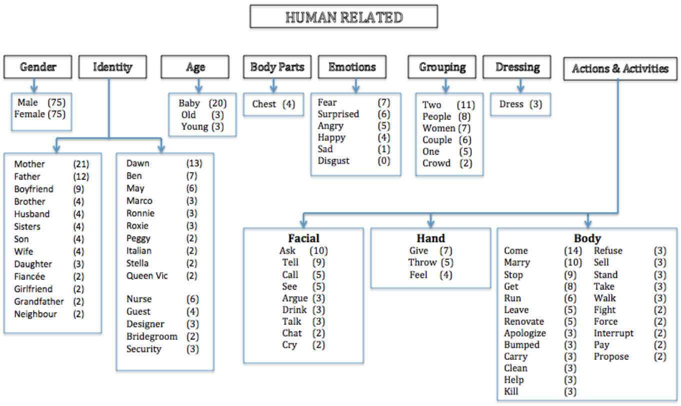


Fig. 7. Human related features in the hand annotations (with audio)

Finding Overlapped Key Interest Points The hand annotations are fed into an automatic summarizer tool[7] to identify the sentence relevance and the best keywords. This automatic summarization tool works in three phases. In the first phase, it will extract the sentences from the input text. Next, it will identify the keywords in the text and count each word's relevance. And in the final phase, it will identify the sentences with the most relevant keywords and displaying them based on the options selected. Table 3 and Table 4 shows the sentences with the highest relevance when the threshold[8] is set to 80. Based on these findings, the overlapped key interest points that have been identified for Video_ID: 5082189274976367100 are: they are two women having a conversation at the hospital; one of the woman ran out from the hospital crying after giving the baby; one of the woman argues with the nurse to get out from the hospital.

Using Audio Data to Reduce Erroneous and Missing Information Table 5 shows the best keywords that were identified. It is shown that when

[7] http://www.tools4noobs.com/summarize

[8] The value used to limit the sentences based on their relevance. The relevance is determined by the number or relevant words in it

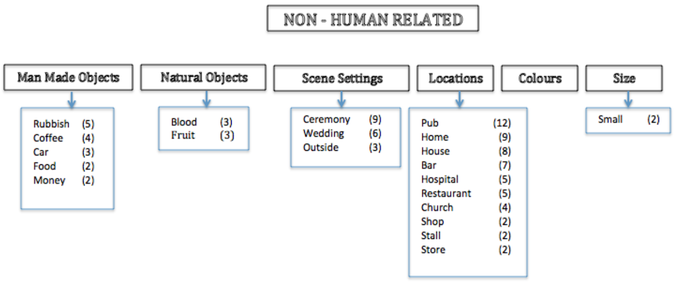


Fig. 8. Non-human related features in the hand annotations (with audio)

Table 3. Sentence relevance for hand annotations without audio (Video_ID : 5082189274976367100)

Summary	**Sentence Relevance**
There is one patient lying at the hospital bed and is talking with another woman who is holding a baby.	40
In the hospital, the young woman prepares to go back while a nurse came to talk to her.	37
One is a young woman sitting on the bed and one is middle age woman standing holding the baby.	32
A woman fought with nurse possibly about getting out from the hospital.	32
She ran out from the hospital and cried after gave the baby to that young lady.	32

audio data is present, the participants are keener towards identifying the identity of the human subjects (e.g., 'Dawn', 'mother'). Besides that, the key information of the video is also identified. In the hand annotations (without audio), although they managed to identify that the two women are having a conversation, the information regarding the conversation itself is missing. Keywords such as 'probably', 'possibly', and 'maybe' were often used. In the hand annotations (with audio), there is a substantial increment in relevance for the keyword 'baby'. This clearly shows that the participants have grasped the key information of the video that is about the two women arguing over the baby. Therefore, we can conclude that incorporating audio data may reduce erroneous and missing information.

This experiment also shows that there are a few challenges to be overcome when these two types of data are incorporated. First is to establish the relation between what is spoken and what is shown visually. The audio information extracted may or may not be related to the events or activities that are happening in that particular scene. For example, a conversation may be something about

Table 4. Sentence relevance for hand annotations with audio (Video_ID : 5082189274976367100)

Summary	Sentence Relevance
The mother begs and cries to get the baby and called the nurse to ask the middle age woman to leave.	45
Dawn was furious and demanded her baby but the lady tried to increase the amount of money.	44
May wants to take the newborn baby from Dawn who is the mother as they agreed before by paying some amount of money.	43
Dawn had just gave birth and a lady was trying to take her baby away and claimed that they had agreed on giving the baby to her in return of GBP10, 000.	41
Dawn want to leave the hospital but the nurse try to stop her because she need some rest.	37
Dawn wants to leave from the hospital with her baby.	37
The baby's mother want to go back home because she worries the middle age woman will come back and steal her baby.	37

Table 5. Best keywords for hand annotations without audio (left) and with audio (right) for Video_ID: 5082189274976367100

Keyword	Relevance
Woman	13
Hospital	10
Baby	8
Nurse	7
Lady	6

Keyword	Relevance
Baby	16
Dawn	11
Stop	6
Mother	6
Nurse	6

the past that differs (non-relevant) from what is visually shown. Future work should consider a decision-making process to filter non-relevant audio information by crosschecking it with the visual features and calculate their overlapped similarities.

Secondly, various audio processing tasks should be incorporated to get optimum results. For example, detecting a person's identity or relationship. Speech recognition alone is not sufficient to determine which spotted keyword can be associated with which detected person. It should include various cues in audio and video to determine either the keyword is referring to the person whom he/she is having the conversation with or a third person that may or may not be present in the video stream. Associating a detected person with a keyword that represents his identity or relationship is a challenge that is yet to be overcome.

Third, this experiment uses the Eastender dataset that has been crafted to include scenes with human activities and events. Thus, it is "rich" in both audio and visual information to highlight the key interest points in the video stream. Different set of guidelines should be given to the participants depending on the type of the video dataset. For example, a lecture video may include a person presenting a PowerPoint slide. Although, the audio features may differ to their detected visual counterpart, in this context the information is relevant to describe the video content. For surveillance videos, the guideline should outline what is expected to be annotated. Due to the nature of this type of dataset that has no clear storyline or video highlights, a clear guideline is crucial to minimize hand annotations that are too diverse or subjective between one another.

Therefore, in order to incorporate audio data towards curbing missing information, these are the challenges that need to be put into consideration to achieve optimum results.

6 Conclusion

This paper has proven that although visual data is sufficient to detect humans, their interactions with related objects, actions, and scenes, using this information alone to generate natural language descriptions may not be able to capture the "key interest point" of the video content. An ideal video summary provides a brief overview of the video. It is not merely stating what is present (detected) in the video. Therefore, incorporating audio data is crucial towards curbing missing or erroneous information. Future work should consider the challenges that may arise when incorporating both of these data primarily the challenge of filtering relevant and non-relevant information. The corpus dataset (hand annotations) can also be used as a mean of evaluation against future works on natural language generation of a video stream.

References

1. Smith, M.A. and Kanade, T.: Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques. In: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference, pp. 775–781. (1997)
2. Lienhart, Rainer and Pfeiffer, Silvia and Effelsberg, Wolfgang: Video Abstracting. In: Commun. ACM, vol. 40, pp. 54–62. New York (1997)
3. He, Liwei and Sanocki, Elizabeth and Gupta, Anoop and Grudin, Jonathan: Auto-summarization of Audio-video Presentations. In: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), pp. 489–498., Orlando, Florida (1999)
4. Uchihashi, Shingo and Foote, Jonathan and Girgensohn, Andreas and Boreczky, John: Video Manga: Generating Semantically Meaningful Video Summaries. In: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), pp. 383–392., Orlando, Florida (1999)
5. Yeung, M.M. and Boon-Lock Yeo: Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content. In: Circuits and Systems for Video Technology, IEEE Transactions, pp. 771–785., (1997)

6. Assfalg, Jürgen and Bertini, Marco and Colombo, Carlo and Bimbo, Alberto Del and Nunziati, Walter: Semantic Annotation of Soccer Videos: Automatic Highlights Identification. In: Comput. Vis. Image Underst., vol. 92, pp. 285–305. New York (2003)
7. Cui, Bin and Pan, Bei and Shen, HengTao and Wang, Ying and Zhang, Ce: Video Annotation System Based on Categorizing and Keyword Labelling. In: Database Systems for Advanced Applications, vol. 5463, pp. 764–767. Springer, Heidelberg (2009)
8. Khan, Muhammad Usman Ghani and Nawab, Rao Muhammad Adeel and Gotoh, Yoshihiko: Natural Language Descriptions of Visual Scenes: Corpus Generation and Analysis. In: Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), pp. 38–47., Avignon, France (2012)
9. Khan, Muhammad Usman Ghani and Lei Zhang and Gotoh, Yoshihiko: Generating coherent natural language annotations for video streams. In: Image Processing (ICIP), 2012 19th IEEE International Conference, pp. 2893-2896, (2012)
10. Niveda Krishnamoorthy and Girish Malkarnenkar and Raymond Mooney and Kate Saenko and Sergio Guadarrama: Generating Natural-Language Video Descriptions Using Text-Mined Knowledge, (2013)
11. Kojima, Atsuhiro and Tamura, Takeshi and Fukunaga, Kunio: Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. In: Int. J. Comput. Vision, vol. 50, pp. 171–184., Hingham, MA, USA (2002)
12. Das, P. and Chenliang Xu and Doell, R.F. and Corso, J.J. : A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference, pp. 2634-2641., (2013)