Video Summarization Using Key-Frame Captions

*Abhijith Aravind, *Alkka Wilson, *Aryan T R, *T G Vidhya, **Sruthy Manmadhan
*UG Scholar, **Assistant Professor, Department of Computer Science and Engineering
NSS College of Engineering Palakkad

NSS COLLEGE OF ENGINEERING
Palakkad, Kerala, India

INTRODUCTION

- Automatic video captioning, where given an input video, a learned model should describe its content in textual format.
- The problem of translating from the visual domain to a textual one is challenging.
- Incorporating the latest technologies in video summarisation, text generation and image captioning, it will be able to make an efficient video summarization system with key-frame captions.
- This is an improved variant among all the key-frame extraction and captioning models as it introduces key-frame extraction and frame captioning modules to be independent of each other.

PROBLEM STATEMENT

- A system that offers a brief semantic understanding of a long video through a text summary.
- To improve the technique of automatic static summary generation from a video using recent technologies.
- To incorporate two complementary tasks (video summarisation and Image captioning) into a single system.

LITERATURE SURVEY

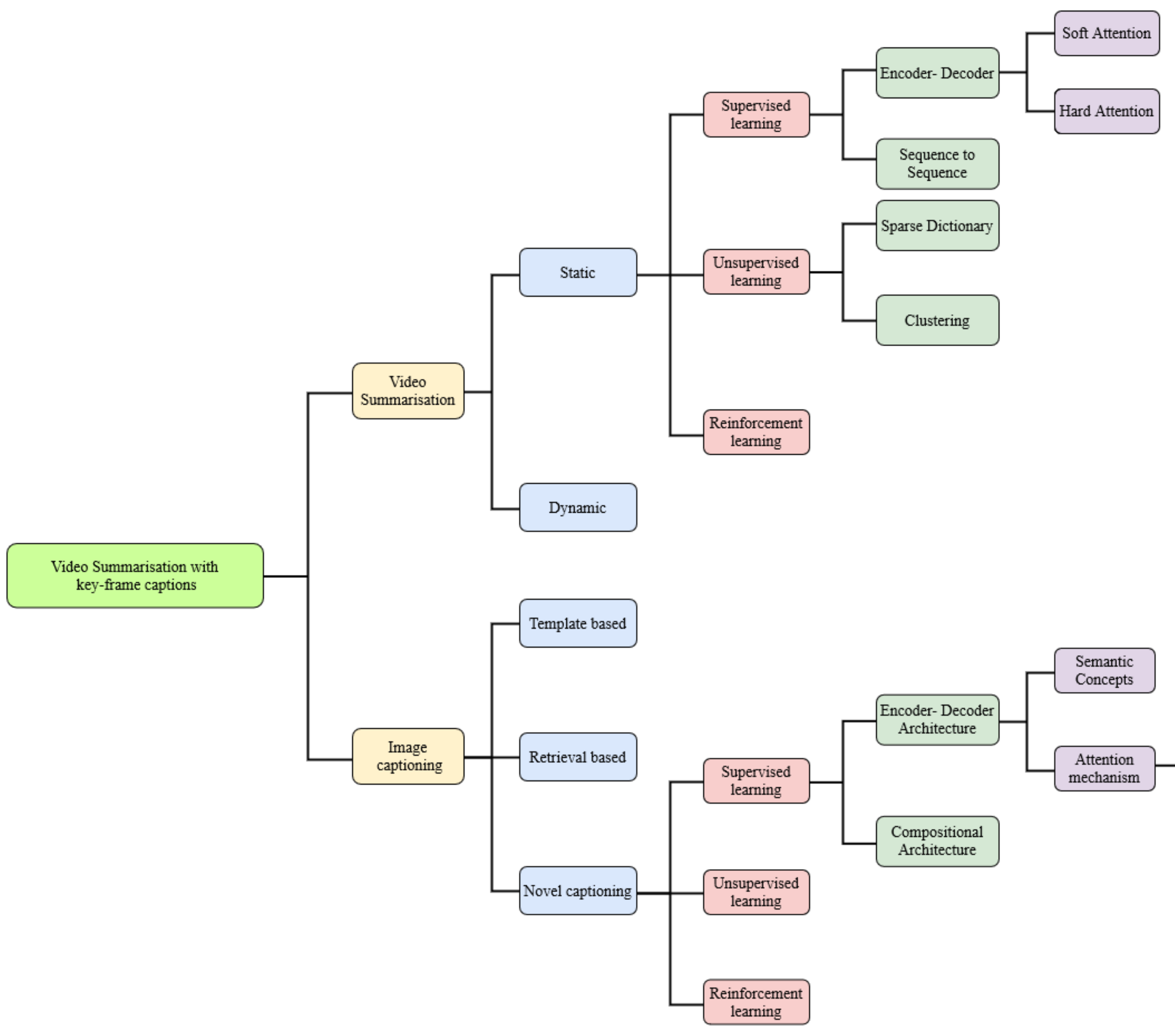

Figure 1: Taxonomy of Video Summarization with Key-Frame Captions

ARCHITECTURE

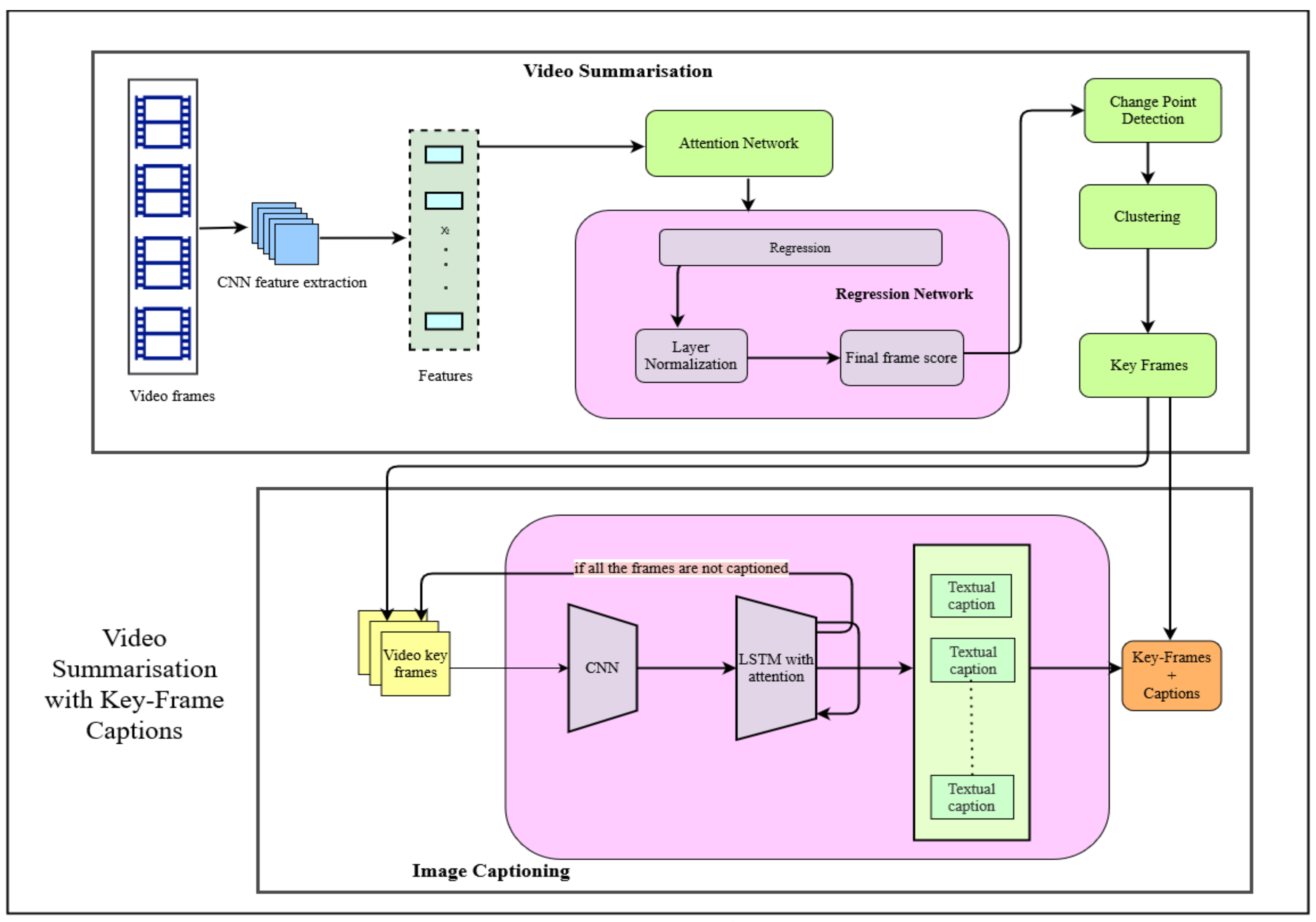

Figure 2: System Architecture

TOOLS & DATASETS

- Datasets - TvSum and SumMe dataset is used for Video Summarization provide frame-level importance score for each video. MSCOCO dataset is used for train Image Captioning model.
- Each annotation in MSCOCO contains category id, image id, caption.

- The system was implemented using tools like Python, Google Colab, InceptionV3, GoogleNet, Flask, Firebase, Heroku, ngrok.
- Following python libraries were used -
  - TensorFlow
  - Keras

Dataset	Videos	User annotations	Annotation type	Video length (sec)		
				Min	Max	Avg
SumMe	25	15-18	keyshots	32	324	146
TvSum	50	20	frame-level importance scores	83	647	235
OVP	50	5	keyframes	46	209	98
YouTube	39	5	keyframes	9	572	196

Figure 3 : Dataset for Key-frame extraction

IMPLEMENTATION STEPS

- A video is given as input and it is sampled to obtain all the frames.
- With the help of a self attention and a fully connected regressor network, each frame will possess a frame level importance score corresponding to them.
- These scores will decide whether the frame should be considered as a key frame or not. Based on the regression scores, the change points are identified and the frames with large scene changes are considered for further filtration.
- These frames are again reduced to fewer key frames by using K-Means clustering.
- The chosen key frames are given to an image captioning with attention mechanism to generate captions corresponding to the frames.
- Beam Search is also used as post processing process for better captions
- Finally the frames and their respective captions are given as the output

RESULT AND ANALYSIS

Video Summarization

Algorithm	Precision	Recall	F-Score
DT	38.06	28.73	31.64
OVP	44.58	50.72	44.81
STIMO	36.65	42.88	37.95
VSUMM	47.44	42.63	43.75
MSR	41.42	53.82	44.98
SOMP	40.92	56.58	45.46
AGDS	38.06	63.48	45.54
SBOMP	42.09	65.65	49.56
PROPOSED SYSTEM	**49.1**	**80.94**	**60.32**

Figure 4 : F-scores of video summarisation.

Image Captioning

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Mao et al. 2015	0.670	0.490	0.350	0.250
Jia et al. 2015	0.670	0.491	0.358	0.264
Fu et al. 2015	0.697	0.519	0.381	0.282
PROPOSED SYSTEM	**0.705**	**0.541**	**0.442**	**0.348**

Figure 5 : BLEU scores of image captioning

**Overall System Evaluation*

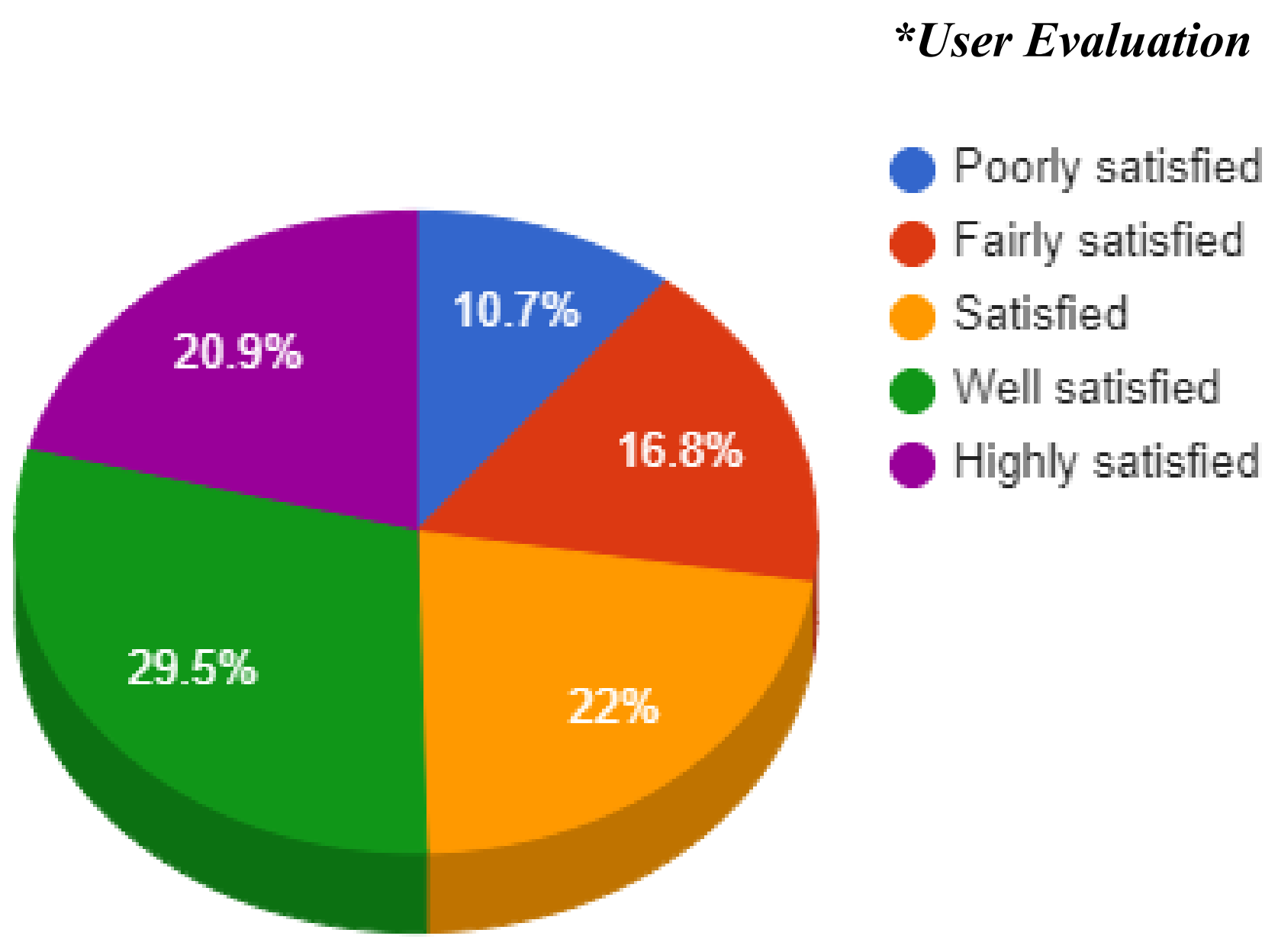

Figure 6 : Graphical representation of user ratings

CONCLUSION

- The results are optimal and perform better than the existing systems.
- Video summarisation model is able to perform even better than all the other models mentioned.
- The image captioning model with attention is equivalent to the state-of-the-art.
- Model is able to moderately satisfy the user's expectations.

FUTURE WORKS

- The model shall be trained on clipart-like objects with another dataset so that it can improve the efficiency of captions.
- The textual captions can be extended to a textual summary using some transformer based algorithm like GPT-2
- Video titles can also be generated based on the overall content of the textual summary

REFERENCE

- Fajtl, Jiri, et al. (2018) Summarizing videos with attention. Asian Conference on Computer Vision. Springer, Cham.
- Xu, Kelvin, et al. (2015) Show, attend and tell: Neural image caption generation with visual attention. International conference on machine learning.
- Kaur, P., and Kumar, R. (2018) Analysis of video summarization techniques. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* 6.1, pp.1157-1162.
- Hani, A., Tagougui, N. and Kherallah, M. (2019) Image Caption Generation Using A Deep Architecture. *2019 International Arab Conference on Information Technology (ACIT)*. IEEE