Multi-Modal Summarization of Read, Watch, Listen for Text and Multimedia Content

P Subhash, Ram Mohan S A

***Abstract*: Conceptual Automatic content summarization is a main NLP apps that intends to consolidate a source content into a shorter adjustment. The quick addition in all kind of data show over the internet requires multi-modal summarization (MMS) from non-simultaneous aggregations of content, picture, sound and video. Here propose an extractive MMS procedure that joins the strategies of NLP, discourse handling and PC vision to examine the rich data contained all kind of data and to get better the idea of multimedia news summarization. The main idea is to associate the semantic openings between multimodel substance. Sound and visual are major modalities in the video. For sound data, we structure an approach to manage explicitly use its interpretation and to find the astounding nature of the translation with sound signals. For visual data, we get acquainted with the joint depictions of content and pictures using a neural framework. By then, we get the incorporation of the made framework for noteworthy visual data through content picture coordinating or multimodal topic showing. Finally, all the multimodal points are considered to make a literary once-over by increasing the striking nature, non-reiteration, clarity and consideration through the arranged streamlining of sub isolated limits.**

***Index Terms*: Summarization, Multimedia, Multi-modal, NLP**

I. INTRODUCTION

Text summarization expects a fundamental employment in our step by step life and has been considered for an extended period of time. With the incident to the data age and the advancement of multimedia development, multimedia data (counting text, picture, audio, video) have extended altogether. Multimedia data have unimaginably changed the way where people live and make it difficult for customers to get noteworthy data capably. Most summarization structures base on just NLP, the opportunity to commonly improve the idea of the diagram with the guide of programmed discourse acknowledgment (ASR) and PC vision (CV) handling systems is commonly dismissed. Normally non-concurrent, all things considered, which means there is no given express depiction for pictures and no subtitles for recordings. Thusly, MMS [1] faces an essential test in understanding the semantics of visual data. Here presented an MMS system that can outfit customers with textual once-overs to acquire the substance of unique multimedia data in a concise range without examining reports or watching recordings all the way.

Conventional record summarization ponders two fundamental edges: (1) Salience: the outline should hold the basic substance of the data archives. (2) Non-abundance: the once-over should contain as small dull substance as would be judicious. For MMS, having two perspectives: (3)Readability: since discourse interpretations are by chance gravely formed, we endeavor to discard the botches displayed by ASR. (4) Coverage for the visual data: pictures that appear in reports and recordings consistently catch event includes that are commonly critical. Along these lines, the once-over should cover anyway a great part of the critical visual data as could be normal.

II. Related Work

Text summarization is to expel the noteworthy data from source archives. With the development of multimedia data on the web, a couple of pros (Shah et al., 2016; Li et al., 2017) revolve around multimodal summarization starting late. Existing examinations (Li et al., 2017, 2018a) have exhibited that, stood out from text summarization; multimodal summarization can improve the idea of delivered abstract by using data in visual modality.

Regardless, the yield of existing multimodal summarization systems is commonly addressed in a single modality, for instance, textual or visual (Li et al., 2017; Evangelopoulos). In this paper, we fight that multimodal output1 is significant for the going with three reasons: 1) It is much easier and faster for customers to get essential data from the pictures (Li et al., 2017). 2) According to our investigations, the multimodal yield (text+image) extends customers' satisfaction by 12:4% stood out from the single-modality yield (text) (more nuances can be found in Sec. 4.2). 3) Images help customers to understand events while

Revised Manuscript Received on May 25, 2019.

Dr. P Subhash, [2]Ram Mohan S A
Associate Professor, M.Tech
Department of Computer Science and Engineering
VNRVJIET, Hyderabad.

texts give more nuances related to the events. As such the pictures and text can enhance each other, helping customers to get an undeniably visualized comprehension of events (Bian et al., 2013). We give a model in Fig. 1 to speak to this wonder. For the yield with simply the text summation, customer will be perplexed about the depiction of "four-legged creatures"; while with an appropriate picture, customer will have an all the more clear comprehension of the text.

Multi-document Summarization

MDS tries to expel huge data from a great deal of reports related to an event to create an outline of much more diminutive size. MDS can be abstractive or extractive. Extractive-based models use distinctive phonetic features, for instance, sentence position [17], [18] and tf*idf [19], to perceive the most surprising sentences in a ton of reports. Diagram based methods [20] are generally used extractive-set up together MDS models based. Finally, the top-situated sentence is picked to manufacture outlines.

Multi-modal Summarization

As of now, much work has been performed to consolidate meeting narratives, sport recordings, films, pictorial storylines and social multimedia. Erol et al. [2] hope to make huge parts of a social event recording subject to an examination of sound, text and visual activity. Tjondronegoro et al. [4] propose a technique to consolidate a game by separating the textual data removed from multiple assets and recognizing the huge substance. Li et al. consolidate news pictures by text and visualize text by pictures.

By then, a news story and a picture are picked to address each topic. For electronic long range interpersonal communication summarization, Fabro et al. [10] and Schinas et al. [12] propose to layout the genuine events subject to multimedia content. A multimodal LDA to recognize topics by getting the connections between's the text and image features of little scale online diaries with introduced pictures. The yield of their procedure is a great deal of specialist pictures that portray the events.

III. PROBLEM DEFINITION

The present applications related to MMS consolidate get-together of summarization, sport video summarization, film summarization, pictorial storyline summarization, course of occasion's summarization and social multimedia summarization. Past examinations on these topics overwhelmingly revolve around sketching out synchronous multimodal substance. Pictorial storylines involve a great deal of pictures with text depictions. None of these applications revolve around sketching out multimedia data that contain non-simultaneous data about events.

IV. IMPLEMENTATION MODEL OVERVIEW

There are various basic edges in making a not too bad textual framework for multi-modal data. The prominent substance in records should be held, and the key convictions in recordings and pictures should be verified. Further, the once-over should be clear and non-dull and should seek after the fixed length prerequisite. All of these perspectives can be as one overhauled by the arranged expansion of sub particular capacities.

$$\text{Max } S\ T\ \{F(S) : X\ s\epsilon S\ ls <= L\ \}$$

Above T is the course of action of sentences, S is the blueprint, ls is length words, L is spending plan, i.e., length prerequisite for the summary, and sub measured limit F(S) is the summation score related to the recently referenced points of view. Text is the essential modality of archives, and on occasion, pictures are embedded in records. Recordings include at any rate two sorts of modalities: sound and visual. Next, we give all things considered handling strategies for different modalities.

Sound, i.e., discourse can be consequently converted into text by using an ASR system2. By then, we can utilize a chart based methodology to figure the striking nature score for most of the discourse translations and for the primary sentences in archives. Note that discourse translations are much of the time ineffectively encircled; as such, to improve the weightiness, we should endeavor to avoid the missteps introduced by ASR.

For visual, which is extremely a progression of pictures (diagrams), in light of the fact that most of the adjacent housings hold abundance data, we first concentrate the most significant edges, i.e., key frames. We become acquainted with the joint depictions for textual and visual modalities and would then have the option to recognize the sentence that is appropriate to the picture. Thusly, we can guarantee the consideration of made framework for the visual data.

Architecture Model

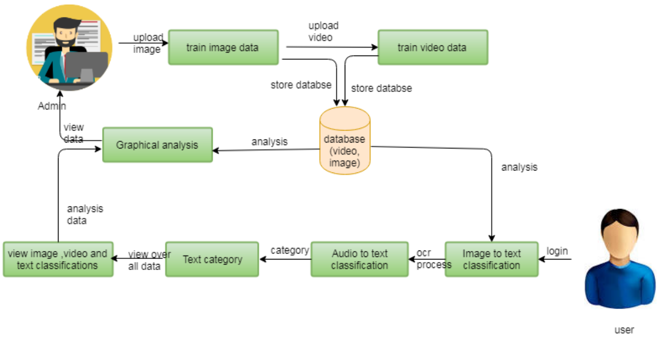


Fig1: Implementation Architecture Work Flow

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

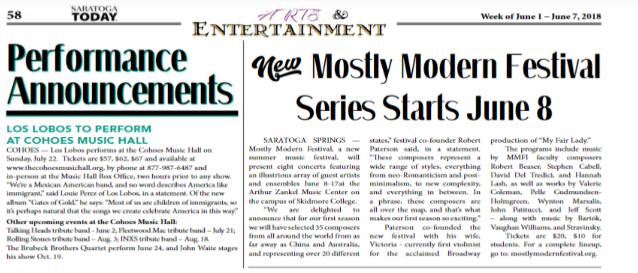
Text-Image Matching Model

The key frames in the recordings and the pictures embedded in the archives as often as possible catch news includes that address the huge data that the blueprint should cover. Before assessing the incorporation for the pictures, we need a model to vanquish any prevention among text and picture. We can deal with this issue by cross-modal examination. Crossmodal semantic coordinating can be better researched when multi-modal data is foreseen into the joint subspace.

Here the text-picture coordinating model is readied, for each text-picture pair (si, pj) in our endeavor, we can discover the coordinating score m(si, pj). We set the edge as the typical coordinating score for the positive text-picture pair.

Semantic Frame Level Text-Image Matching

The basic idea is that each activity word in a sentence is set apart with its propositional disputes, and the naming for each particular activity word is known as a "plot". Each packaging addresses an event, and the conflicts express the significant data about this event. There is a great deal of disputes showing the semantic activity of each term in a packaging. An instance of edge semantic parsing is exhibited wail figure. The main sentence "President Bush affirmed government calamity help for the affected domains and made courses of action for an examination voyage through the state" is changed into two modified sentences "President Bush endorsed bureaucratic disaster help" and "President Bush made game plans for an examination voyage through the state". The improved sentences have less tolerable assortment in significance, which focal points the text-picture coordinating.



58 SARATOGA TODAY ARTS & ENTERTAINMENT Week of June 1 – June 7, 2018

Performance Announcements

LOS LOBOS TO PERFORM AT COHOES MUSIC HALL

COHOES — Los Lobos performs at the Cohoes Music Hall on Sunday, July 22. Tickets are $57, $62, $67 and available at www.thecohoesmusichall.org, by phone at 877-987-6487 and in-person at the Music Hall Box Office, two hours prior to any show. "We're a Mexican American band, and no word describes America like immigrant," said Louie Perez of Los Lobos, in a statement. Of the new album "Gates of Gold," he says: "Most of us are children of immigrants, so it's perhaps natural that the songs we create celebrate America in this way."

Other upcoming events at the Cohoes Music Hall:
Talking Heads tribute band - June 2; Fleetwood Mac tribute band – July 21; Rolling Stones tribute band – Aug. 3; INXS tribute band – Aug. 18.
The Brubeck Brothers Quartet perform June 24, and John Waite stages his show Oct. 19.

New Mostly Modern Festival Series Starts June 8

SARATOGA SPRINGS — Mostly Modern Festival, a new summer music festival, will present eight concerts featuring an illustrious array of guest artists and ensembles June 8-17at the Arthur Zankel Music Center on the campus of Skidmore College.

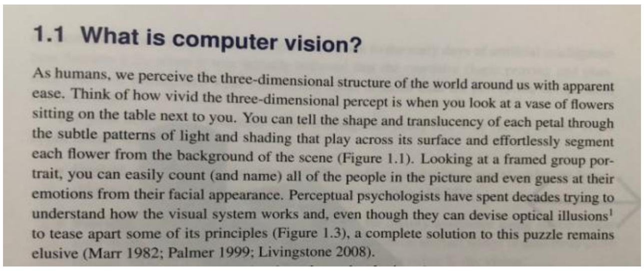
"We are delighted to announce that for our first season we will have selected 35 composers from all around the world from as far away as China and Australia, and representing over 20 different states," festival co-founder Robert Paterson said, in a statement. "These composers represent a wide range of styles, everything from neo-Romanticism and post-minimalism, to new complexity, and everything in between. In a phrase, these composers are all over the map, and that's what makes our first season so exciting."

Paterson co-founded the new festival with his wife, Victoria - currently first violinist for the acclaimed Broadway production of "My Fair Lady."

The programs include music by MMFI faculty composers Robert Beaser, Stephen Cabell, David Del Tredici, and Hannah Lash, as well as works by Valerie Coleman, Pelle Gudmundsen-Holmgreen, Wynton Marsalis, John Patitucci, and Jeff Scott – along with music by Bartók, Vaughan Williams, and Stravinsky.

Tickets are $20, $10 for students. For a complete lineup, go to: mostlymodernfestival.org.

Fig2: Image and the potential matched text in our dataset.



1.1 What is computer vision?

As humans, we perceive the three-dimensional structure of the world around us with apparent ease. Think of how vivid the three-dimensional percept is when you look at a vase of flowers sitting on the table next to you. You can tell the shape and translucency of each petal through the subtle patterns of light and shading that play across its surface and effortlessly segment each flower from the background of the scene (Figure 1.1). Looking at a framed group portrait, you can easily count (and name) all of the people in the picture and even guess at their emotions from their facial appearance. Perceptual psychologists have spent decades trying to understand how the visual system works and, even though they can devise optical illusions[1] to tease apart some of its principles (Figure 1.3), a complete solution to this puzzle remains elusive (Marr 1982; Palmer 1999; Livingstone 2008).

Fig3: Example for simplified sentence based on frame-semantic parsing.

Chunking Level Text-Image Matching

Piecing, which is the route toward section a long sentence into syntactic non-covering constituents, i.e., phrases, is a run of the mill framework in NLP. The advantage of lumping is that the isolated expressions are semantic and noteworthy. Bumps, especially noun phrases (NPs) and action noun phrases (VPs), give a sensible component of reflection over normal language. An action word and other related fragments, including modal, colleague, and modifier. Above figure shows that NPs and VPs give a clear strategy to arrange text and picture.

Multi-modal Topic Modeling

After the text-picture coordinating model is readied, we secure the joint depiction of text and pictures. Next, we perceive the topics of text and pictures. The motivation driving this method is that textual delineations of pictures as often as possible give noteworthy data about semantic edges (topics), and picture features are consistently connected with semantic topics. Wang et al. make a course of occasions summarization for Tweet streams by recognizing topic advancement. For our errand, the multi-modal topic model can reveal various pieces of text and pictures; by then we can explore a representative arrangement of text covering the pieces of the pictures. Topic models, for instance, LDA, can together learn inert topics and topic assignments of reports. To reveal the semantic perspectives, we make the multi-modal topic model subject to a neural topic model (NTM). The multimodal topic model figures the unforeseen probability p(w|d) using the apportionment of the word (or picture)- topic p(w|t) and topic-report (or video) p(t|d).

$$p(w|d) = \sum^{T}{}_{i=1}\ p(w \mid ti)p(ti \mid d)$$

V. MULTI-MODAL SUMMARIZATION

Text Salience of Summarization

The game plan of the significant number of sentences and discourse interpretations V into autonomous packs, and the striking nature scores s(tj) are institutionalized to [0,1] by isolating by the most outrageous motivating force among all of the sentences. This objective limit prizes striking nature and not too bad assortment since it is progressively helpful to pick a sentence from a gathering that of its parts in the once-over. If a sentence is investigated a gathering, various sentences from this bundle have a reducing increment as a result of the square root work.

Matching-based Image Coverage of Summarization

For key frame pi, Im (pi) is the ordinary striking nature score of the discourse translations inside the shot to which pi has a spot. For archive picture pi, Im(pi) is the typical striking nature score of the sentences

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

in the record in which pi is introduced. cj is a sentence or a sentence segment got reliant on semantic enclosing, piecing or word tokenizing.

This objective limit hopes to grow the weighted consideration of the picked pictures. For a picture pi, only a cj with the most extraordinary coordinating score with pi adds to the consideration for pi. That is, when pi is verified by cj , no other ck can further improve the incorporation of pi. The sense behind this is immediate: the delivered summation is limited long, and we intend to cover anyway numerous critical pictures as would be reasonable. Here consider not simply the count of the pictures verified by the framework yet what's more the general centrality score of the verified pictures. In this manner, we select sentences that are material to dynamically critical pictures or that contain more parts relevant to progressively huge pictures.

Topic-based Image Coverage of Summarization

This objective limit hopes to enhance the weighted topic consideration of the picked pictures. For a specific pi, such as coordinating based picture incorporation, only a sentence or discourse translation sj with the most outrageous topic-based comparability score with pi adds to the consideration for pi, and no other sk can further improve the incorporation of pi. Appropriately, the created diagram covers whatever number critical topics of the pictures as would be judicious.

VI. DATA COLLECTION AND ANNOTATION

Here build up a dataset as seeks after. We select haphazardly underground creepy crawly event related from the most recent five years English figuratively speaking. Here assembled 10 records, pictures. We use 10 graduate understudies to make reference outlines in the wake of scrutinizing records and watching recordings on a comparative event. The criteria for gathering records are (1) hold the huge substance of the information reports (2) avoid monotonous data; (3) have a respectable lucidity; (4) satisfy quite far.

VII. EXPERIMENTAL STUDIES

A couple of models are pondered in our examinations, incorporating delivering once-overs with different modalities and using different approaches to manage impact pictures.

Text only This model produces diagrams using just the text in archives.

Audio only This model produces blueprints using just the discourse translations from recordings.

The going with models produce layouts using the two archives and recordings anyway adventure pictures in different ways. The striking quality scores for text are gained with heading techniques.

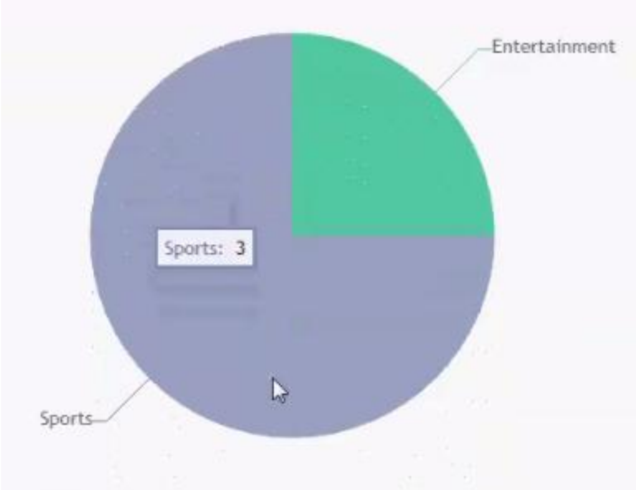


Fig4: Audio Analysis

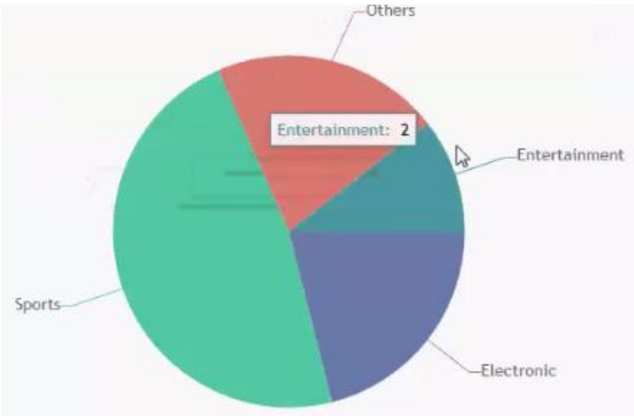


Fig5: Image Analysis

VIII. CONCLUSION

This paper tends to an offbeat MMS task, to be specific, how to utilize related text, sound and video data to create a textual outline. We define the MMS task as an advancement issue with a planned expansion of sub measured capacities. We address comprehensibility by specifically utilizing the translation of sound through direction methodologies. All the more explicitly, we plan a novel chart based model to successfully figure the notability score for every text unit, prompting progressively meaningful and useful rundowns. We examine different ways to deal with distinguish the pertinence between the picture and text.

REFERENCES

[1] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video." in EMNLP, 2017, pp. 1092–1102.

[2] B. Erol, D.-S. Lee, and J. Hull, "Multimodal summarization of meeting recordings," in ICME, vol. 3. IEEE, 2003, pp. III–25.

[3] R. Gross, M. Bett, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Towards a multimodal meeting record," in ICME, vol. 3. IEEE, 2000, pp. 1593–1596.

[4] D. Tjondronegoro, X. Tao, J. Sasongko, and C. H. Lau, "Multimodal

summarization of key events and top players in sports tournament videos," in WACV. IEEE, 2011, pp. 471–478.

[5] T. Hasan, H. Boˇril, A. Sangwan, and J. H. Hansen, "Multi-modal highlight generation for sports videos using an informationtheoretic excitability measure," EURASIP Journal on Advances in Signal Processing, vol. 2013, no. 1, p. 173, 2013.

[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," IEEE Transactions on Multimedia, vol. 15, no. 7, pp. 1553–1568, 2013.

[7] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multimodal stereoscopic movie summarization conforming to narrative characteristics," IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5828–5840, 2016.

[8] D. Wang, T. Li, and M. Ogihara, "Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs." in AAAI, 2012.

[9] W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in NAACL-HLT, 2016, pp. 58–68.

[10] M. Del Fabro, A. Sobe, and L. B¨osz¨ormenyi, "Summarization of real-life events based on community-contributed content," in The Fourth International Conferences on Advances in Multimedia, 2012, pp. 119–126.

[11] J. Bian, Y. Yang, and T.-S. Chua, "Multimedia summarization for trending topics in microblogs," in CIKM. ACM, 2013, pp. 1807– 1812.

[12] M. Schinas, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, and P. A. Mitkas, "Multimodal graph-based event detection and summarization in social media streams," in Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015, pp. 189–192.

[13] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua, "Multimedia summarization for social events in microblog stream," IEEE Transactions on Multimedia, vol. 17, no. 2, pp. 216–228, 2015.

[14] R. R. Shah, A. D. Shaikh, Y. Yu, W. Geng, R. Zimmermann, and G.Wu, "Eventbuilder: Real-time multimedia event summarization by visualizing social media," in Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015, pp. 185–188.

[15] R. R. Shah, Y. Yu, A. Verma, S. Tang, A. D. Shaikh, and R. Zimmermann, "Leveraging multimodal information for event summarization and concept-level sentiment analysis," Knowledge-Based Systems, vol. 108, pp. 102–109, 2016.

[16] S. Khuller, A. Moss, and J. S. Naor, "The budgeted maximum coverage problem," Information Processing Letters, vol. 70, no. 1, pp. 39–45, 1999.

[17] V. Varma, V. Varma, and V. Varma, "Sentence position revisited: a robust light-weight update summarization 'baseline' algorithm," in International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, 2009, pp. 46–52.

[18] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, "A study on position information in document summarization," in COLING, 2010, pp. 919–927.

[19] D. R. Radev, H. Jing, M. Sty´s, and D. Tam, "Centroid-based summarization of multiple documents," Information Processing Management, vol. 40, no. 6, pp. 919–938, 2004.

[20] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in ACL, 2004.

[21] X. Wan and J. Yang, "Improved affinity graph based multidocument summarization," in NAACL, 2006, pp. 181–184.

[22] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," Journal of Qiqihar Junior Teachers College, vol. 22, p. 2004, 2011.

[23] X. Zhou, X. Wan, and J. Xiao, "Cminer: Opinion extraction and summarization for chinese microblogs," IEEE Transactions on Knowledge & Data Engineering, vol. 28, no. 7, pp. 1650–1663, 2016.

[24] X. Li, L. Du, and Y. D. Shen, "Update summarization via graphbased sentence ranking," IEEE Transactions on Knowledge & Data Engineering, vol. 25, no. 5, pp. 1162–1174, 2013.

[25] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video." in EMNLP, 2017, pp. 1092–1102.

[26] B. Erol, D.-S. Lee, and J. Hull, "Multimodal summarization of meeting recordings," in ICME, vol. 3. IEEE, 2003, pp. III–25.

[27] R. Gross, M. Bett, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Towards a multimodal meeting record," in ICME, vol. 3. IEEE, 2000, pp. 1593–1596.

[28] D. Tjondronegoro, X. Tao, J. Sasongko, and C. H. Lau, "Multimodal summarization of key events and top players in sports tournament videos," in WACV. IEEE, 2011, pp. 471–478.

[29] T. Hasan, H. Boˇril, A. Sangwan, and J. H. Hansen, "Multi-modal highlight generation for sports videos using an informationtheoretic excitability measure," EURASIP Journal on Advances in Signal Processing, vol. 2013, no. 1, p. 173, 2013.

[30] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," IEEE Transactions on Multimedia, vol. 15, no. 7, pp. 1553–1568, 2013.

[31] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multimodal stereoscopic movie summarization conforming to narrative characteristics," IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5828–5840, 2016.