International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Video Summarization: Survey

Suraj Fule[1], Neel Chhabada[2], Himanshu Choudhary[3], Aishwary Dubey[4]
[1, 2, 3, 4]*Department of Computer Engineering, MIT College of Engineering, Pune, Maharashtra*

Abstract— Sound is one of human beings most important senses. After vision it is the sense most used to gather the information about the environment. Sound information in videos plays an important role in shaping the user feelings and experience. When sound is not available in videos, text captions are used to provide sound information. However, standard text captions are not very expressive and efficient for non-verbal sounds because they are specifically designed to visualize speech sounds.

Keywords—Video Summarization, Data Preprocessing, Feature Extraction, Classification, Sound visualization

I. INTRODUCTION

Recognizing and understanding context of sounds in surrounding environment is very important in the terms of making the next move based on a sound that occurred. Visualization is a tool for both exploration and communication. Since interactive visualization is the main key for intelligently exploration, animation and effectively conveys the complex process or structure. Animation contributes a powerful means for illustrating objects, evolution and interaction in a complex environment. Most visualization systems provide some animation support.

However, text captions are normally static text at the bottom of the screen, which is not very effective for describing non-verbal sounds, such as those from the environment or sound effects (e.g. engine noise in car-racing video). Further, the dynamics of sounds, such as intensification or attenuation, is important for describing non-verbal sounds, but conveying such dynamics is extremely difficult. Moreover, when multiple sounds are mixed together, users have difficulty in identifying where a sound comes from with static captions.

Researches aiming at converting temporal images of videos to comic style images have been reported where methods for showing text near the speaker on images were proposed. However, these methods mainly focused on subtitles of spoken language. In addition, since they add text to still images, the dynamics of sounds are not fully presented. In contrast, our framework adds sound words to videos and can represent the dynamics of sounds by means of animation. The automatic annotation concept we present enables us to easily obtain a video annotated with sound words to enhance the video-watching experience. Such a concept has not been well explored yet.

First, the sound processing step identifies sounds throughout the video and computes the time-series posterior probability for each predefined sound category (e.g. engine sound). The algorithm therefore estimates the sound volume for each pre-defined sound category. Second, the animation generation step uses the result of volume estimation for each sound category to generate the sound-word animation items. Finally, the generated animation items are positioned in the video in a way that indicates the position of the sound source object in the video frames.

The aim to design a system that can notify a user of environmental sound without he or she listening. Furthermore, the user can avoid the dangers through the detection and notification of danger sounds. In this section, we will first introduce the construction of our system, and then we will explain how to recognize a sound source. Finally, we will introduce the sound source visualization system.

II. LITERATURE REVIEW

“Chong-Wah Ngo”, et.al introduces unified approach for summarization based on the analysis of video structures and video highlights. Approach emphasizes both the content balance and perceptual quality of a summary. Normalized cut algorithms are employed to globally and optimally partition a video into clusters. A motion attention model based on human perception is employed to compute the perceptual quality of shots and clusters. The clusters, together with the computed attention values, form a temporal graph similar to Markov chain that inherently describes the evolution and perceptual importance of video clusters.

“Junwei Han,” et al describes Meaningful representation and effective retrieval of video shots in a large-scale database has been a profound challenge for the image/video processing and computer vision communities. A great deal of effort has been devoted to the extraction of low-level visual features, such as color, shape, texture, and motion for characterizing and retrieving video shots. However, the accuracy of these feature descriptors is still far from satisfaction due to the well-known semantic gap. In order to alleviate the problem, this paper investigates a novel methodology of representing and retrieving video shots using human-centric high-level features derived in brain imaging space (BIS) where brain responses to natural stimulus of video watching can be explored and interpreted. At first, our recently developed dense individualized and common connectivity-based

cortical landmarks (DICCCOL) system is employed to locate large scale functional brain networks and their regions of interests (ROIs) that are involved in the comprehension of video stimulus.

"Rushil Anirudh" , et.al focuses Many applications benefit from sampling algorithms where a small number of well-chosen samples are used to generalize different properties of a large dataset. In this paper, we use diverse sampling for streaming video summarization. Several emerging applications support streaming video, but existing summarization algorithms need access to the entire video which requires a lot of memory and computational power. A memory efficient and computationally fast, online algorithm that uses competitive learning for diverse sampling. Algorithm is a generalization of online K-means such that the cost function reduces clustering error, while also ensuring a diverse set of samples. The diversity is measured as the volume of a convex hull around the samples.

"Juhi Naik" et al describes tool to generate a summary of the most salient parts of videos. Unlike most research going on in the field of video compression, instead of decreasing redundancy, we try to shorten the video by skipping the "uninteresting" parts. A new approach has been tried for scoring importance of frames. We try 2 models, Convolutional Neural Nets (CNNs) and CNNs combined with Long Short-Term Memory (LSTM) modules and find that the latter works much better on video data. A different cost function was also tried, using Kullback-Leibler divergence to solve the regression problem instead of MSE.

"Adway Mitra,", et.al A video is understood by users in terms of entities present in it. Entity Discovery is the task of building appearance model for each entity (e.g. a person), and finding all its occurrences in the video. We represent a video as a sequence of trackless, each spanning 10-20 frames, and associated with one entity. We pose Entity Discovery as track let clustering, and approach it by leveraging Temporal Coherence (TC): the property that temporally neighboring tracklets are likely to be associated with the same entity. Our major contributions are the first Bayesian nonparametric models for TC at tracklet-level. We extend Chinese Restaurant Process (CRP) to TCCRP, and further to Temporally Coherent Chinese Restaurant Franchise (TC-CRF) to jointly model entities and temporal segments using mixture components and sparse distributions. For discovering persons in TV serial videos without meta-data like scripts, these methods show considerable improvement over state-of-the-art approaches to tracklet clustering in terms of clustering accuracy, cluster purity and entity coverage.

"Yifang Yin" et al Presents automatic video summary generation with personal adaptations. The user interests are mined from their personal image collections. To reduce the semantic gap, we propose to extract visual representations based on a novel semantic tree (SeTree). A SeTree is a hierarchy that captures the conceptual relationships between the visual scenes in a codebook. This idea builds upon the observation that such semantic connections among the elements have been overlooked in prior work. To construct the SeTree, we adopt a normalized graph cut clustering algorithm by conjunctively exploiting visual features, textual information and social user image connections. By using this technique, we obtain 8.1% improvement of normalized Discounted Cumulative Gain (nDCG) in personalized video segments ranking compared with existing methods. Furthermore, to promote the interesting parts of a video, we extract a space-time saliency map and estimate the attractiveness of segments by kernel fitting and matching.

"Sinnu Susan Thomas" et al The enormous growth of video content in recent times has brought up the need to abbreviate the content for human consumption. There is thus a need for summaries of a quality that comes up to the requirements of human users. This also means that summarization must incorporate the peculiar features of human perception. A new framework for video summarization. Unlike many available summarization algorithms which utilize only statistical redundancy, the first time features of the human visual system (HVS) within the summarization framework itself to allow for the emphasis of perceptually significant events while simultaneously eliminating perceptual redundancy from the summaries.

"Ana Garcia Del Molino.", et.al the introduction of wearable video cameras (e.g. GoPro) in the consumer market has promoted video life-logging, motivating users to generate large amounts of video data. This increasing flow of first-person video has led to a growing need for automatic video summarization adapted to the characteristics and applications of egocentric video. With this paper, the first comprehensive survey of the techniques used specifically to summarize egocentric videos. We present a framework for first-person view summarization and compare the segmentation methods and selection algorithms used by the related work in the literature.

III. EXISTING SYSTEM

The Sound Recognition Process

The most recognition and classification problems are implemented using three-stage process:

Data Preprocessing
Feature Extraction

Classification

The sequence of these three stages is shown in following figure,

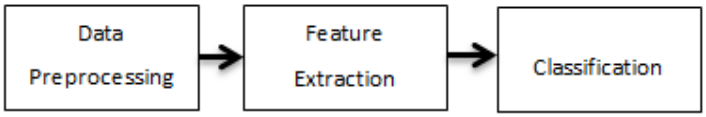

Figure: - Traditional Classification Sequence.

A. Data Preprocessing

It is the first step in the process. This step is dependent on the classification of tasks that are performed. Data Preprocessing for sound recognition involves taking the sound from environment and load it into a computer. Typically it is done using microphone.

In additional computer represent sound in digital format, which means that the analog signals produced by the microphone has to converted into digital format via sampling and quantization technique.

B. Feature Extraction

It is then performed to reduce the huge data set produced in the previous step. It involves selecting pieces of data that uniquely characterize that information. For sound recognition many techniques has been used for feature extraction from the simple to the extravagant. Feature Extraction can be performed at three levels of understanding, All these three levels of understanding can be combine together to produce a system that perform good Feature Extraction.

TABLE I
LEVELS OF FEATURE EXTRACTION

1. Statistical Feature Extraction	Data Based
2. Syntactical Feature Extraction	Data with structure
3. Semantic Feature Extraction	Prior knowledge of Environment

C. Classification

The third step in recognition process. Classification involves taking the feature generated in the previous step and linking each feature to a particular classification. For sound recognition, many techniques are used including hidden Markova models, neural network and Reference model database. All these techniques used training and testing paradigm. Training gives the system a series of particular items, so it can learn the general characteristics of that item. Testing is performed so that it can identify the class of the item that is being tested. As an optional step classification can also done using Fuzzy logic processing.

IV. ADVANTAGES

A. Enhances and Improve visual communication.

B. Its helps to make learning process more effective and conceptual.

C. It provides a realistic approach and experience.

V. APPLICATIONS

Video Summarization.

VI. CONCLUSION

We presented a framework for automatically recognizing non-verbal sounds in video and visualizing the sounds with sound words that are animated based on the volume of the sound it represents. Within this framework, we also presented dynamical positioning the generated sound-word animation depending on the position of the sound source object. We implemented an experimental system to conduct a user study to investigate the effects of the presented sound-word visualization scheme on the user experience.

The result of the user study showed that animated sound words could effectively and naturally visualize the dynamics of sound. This is useful when video is watched without sound and contributes to making watching videos enjoyable. These dynamic positioning methods can clarify the position of a sound source and increase the visual impact.

REFERENCES

[1] Sam Ferguson, Andrew Vande Moere and Densil Cabrera" Seeing Sound : Real-time Sound Visualisation in Visual Feedback Loops used for Training Musicians"
[2] Raisa Rashid, Jonathan Aitken, and Deborah I. Fels" Expressing Emotions Using Animated Text Captions" Springer-Verlag Berlin Heidelberg 2006.
[3] Hiroshi Akiba, Chaoli Wang, and Kwan-Liu Ma _ University of California, Davis" AniViz: A Template-Based Animation Tool for Volume Visualization" 2010 IEEE.
[4] Huadong Wu, Mel Siegel, Pradeep Khosla" Vehicle Sound Signature Recognition by Frequency Vector Principal Component Analysis" May18-20,1998.
[5] Guangmei Jing, Yongtao Hu, Yanwen Guo, Yizhou Yu, Wenping Wang" Content-Aware Video2Comics with Manga-Style Layout" 2015 IEEE.
[6] Angelos Pillos , Khalid Alghamidi , Noura Alzamel , Veselin Pavlov,Swetha Machanavajhala" A REAL-TIME ENVIRONMENTAL SOUND RECOGNITION SYSTEM FOR THE ANDROID OS" 3 September 2016, Budapest, Hungary.
[7] Ruiwei SHEN, Tsutomu TERADA, Masahiko TSUKAMOTO" A System for Visualizing Sound Source using Augmented Reality" MoMM December 3 - 5, 2012.
[8] Kazushi Ishihara,Yuya Hattori,Tomohiro Nakatani,Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno," Disambiguation in Determining Phonemes of Sound-Imitation Words for Environmental Sound Recognition".