International Journal of Electronics Engineering Research.
ISSN 0975-6450 Volume 9, Number 1 (2017) pp. 15-27

http://www.ripublication.com

Utilizing Image Classification based Semantic Recognition for Personalized Video Summarization

Priyamvada R Sachan [#1] and Keshaveni N[#2]

[#1] ECE Department, Visvesvaraya Technological University, MVJ College of Engineering, Bangalore, India.

[#2] ECE Department, Visvesvaraya Technological University, K V G College of Engineering, Sullia, India.

Abstract

In recent past, there had been many attempts to solve the problem of Video Summarization. Most of the state-of-the-art techniques have approached the problem as compression of the information content of video. Such techniques primarily focus on identifying and eliminating redundant frames while attempting to retain all independent concepts in the video. All such techniques often produce single summary for a video. However, when we look from users' perspective, summary of a video can be defined in varying ways as different users may give differential weightage to different independent concepts present in video. Here, we present a novel approach to generate semantic personalized summaries of video for multiple users.

Keyword: Concepts, Personalized Summary, Semantic, Compression.

INTRODUCTION

The alarming growth in videos over the internet has created need for sophisticated and scalable systems for managing the video content for effective user consumption. Busy user demands intelligent techniques for video organization and presentation. As the amount of video content at a user disposal is increasing and effective time to consume all is decreasing, it demands ways to present the content of user's interest in most succinct manner.

Among many other techniques, researchers and video providers have attempted to manage the video problem with video summarization. *Video Summarization* can be defined as condensed representation of a video delivering important and maximum information of the same [1], [2].

As the user base is increasing, inherent diversity among users is also increasing. A single summary for all users does not fit the need. Users depending upon their needs, interests and time-to-consume may consider specific sections of video as important that could be different from other users. This necessitates shift of focus from generic summarization to personalized video summarization.

In this work, we present a generic system to generate personalized video summaries based on semantic aggregation leveraging taxonomy of concepts and explicit modeling of user interests. Section II talks about related research followed by section III, which presents different viewpoints towards video summaries, section IV presents our proposed methodology, section V highlights entity identification using image classification, section VI talks about user model representation and experimental results & conclusion are presented in sections VII & VIII

BACKGROUND

In recent years, there has been growing inclination towards employing deep semantic techniques for solving the general problem of video content management. In this regard, image classification has gained increasing popularity in leading research areas like content-based image retrieval, object recognition and has also made its impact into video summarization. It relies primarily on Bag-of-features (BOF) model, which draws ideas from text/document classification's Bag-of-words (BOW) technique [3], [4]. It essentially incorporates three main steps to classify the image/text: Feature extraction, Quantization and Classification. Image classification was successfully adopted in automated animal species identification [5]. This work was based on scale invariant feature transform (SIFT) and local binary patterns (LBP) for local feature extraction over which spatial pyramid matching is performed. The advantage of this technique is that it also captures the spatial information of the keypoints, which makes the matching process more efficient.

Another popular approach for extraction of semantic information from a video image is based on extraction of semantics from its subtitles through natural language processing (NLP) [6].

As semantic techniques have helped to gain better control in video summarization, focus has shifted towards more personalized video summarization, which is intended to present the user with essential core concise information from a video according to his/her interests.

In [7] authors have proposed personalized video summary generation based on user's estimates such as: like/dislike/neutral. The estimated video segments (scenes) are then used to detect the objects whose importance for user is calculated based on probability estimates. All such objects having impact on user's estimation are referred as user's

range of interest (ROI) and are considered in summary. Inclusion of user model thus makes the summaries more concise and befitting to user's interest.

DIFFERENT VIEWS OF A VIDEO SUMMARY

A video is a communication medium comprising of a sequence of visual frames carrying some information intended by creator of the content. These informative frames present multiple distinct concepts, however, many of the other video frames can be considered as just elaborations of some already presented concepts. One view of summary could be to eliminate many of these informative elaborative frames while retaining all the distinct concepts present in the original video. This approach can be termed as *'compression'* based summarization.

When a user watches the same video, he/she may gather only a concise perspective of all information present, which is often a subset of original concepts induced by creator. This is due to differences in user & creator's preferences about the concepts mentioned in the video. On the similar lines, different users with varying preferences may gather different perspectives of video from each other.

We consider the intent of video summarization to be able to present those frames which makes most sense to a user based on his/her interests. We term this as *'selection'* based summarization. The paper proposes to leverage a user-specific model for ranking and selecting concepts deemed important by user from video for generating user specific summary.

PROPOSED METHODOLOGY: VIDEO SUMMARY AS PERSONALIZED SEMANTIC AGGREGATION OF CONCEPTS

The novel video summarization mechanism presented here attempts to combine the benefits of both compression and selection based approaches. The overall process is divided into multiple steps. Fig.1 represents the overall block diagram of the proposed summarization pipeline. Different steps of the proposed summarization pipeline are detailed below:

Step 1: Segregation of video into Framesets

Original input video is broken into its atomic frames and thereafter contiguous frames that are semantically similar are joined together to create framesets. A *frameset* here is defined as a group of contiguous frames representing same visual concept. A frame similarity function ($Sim_{frame)}$ is used for determining similarity between two frames. Candidates for frame similarity function are based on image features like color, edge and combination of both. It was found that weighted combination of both color and edge performs better as cited in our previous work [8]. An ungrouped frame starts a new frameset and successive frames can be added to the existing frameset iff the following conditions are satisfied:

(i) Similarity of the candidate frame with first frame of frameset is above a ***sim_threshold***.

(ii) Frameset size is within ***frameset_upperbound***.

If a candidate frame cannot be added to existing frameset, it starts a new frameset. *sim_threshold* & *frameset_upperbound* are appropriately chosen to maximize chances of generated framesets consisting of frames of single visual concept. Frameset generation is presented in detail in our previous work [8].

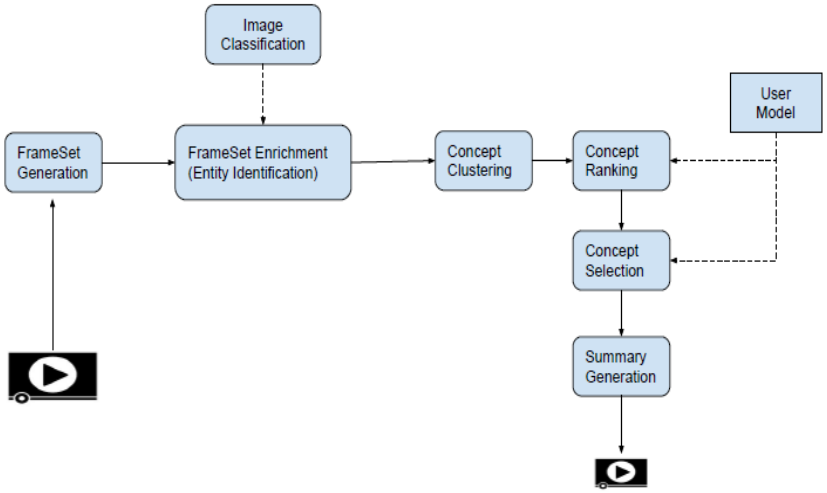


Fig. 1. Block Diagram of Proposed Summarization Pipeline

Step 2: Frameset Enrichment (Concept Identification)

This step attempts to enrich the framesets by explicitly identifying their embedded concept by applying *Image Classification* on its first frame. Since, all frames of a frameset are supposed to represent same concept, classification of its first frame is expected to correctly represent frameset concept. Classification of the first frame is done against a hierarchical taxonomy of domain specific concepts. Details of image classification system and hierarchical taxonomy are presented in section V.

Step 3: Concept Clustering

This step attempts to cluster together different framesets from video that are semantically close to each other to form *concept-clusters*. These concept-clusters are intended to represent macro-visual concepts created from similar micro visual concepts from different sections of video. Also, Framesets are grouped using agglomerative (bottom-up) clustering technique based on following pairwise similarity function (Sim_{FS}) on two framesets as in Eq. 1. This pairwise similarity

function is defined as a weighted sum of category-similarity and frame similarity between two framesets, where w_{cat} represents the weightage of category-similarity in frameset similarity:

$$Sim_{FS}(fsa, fsb) = w_{cat} * Sim_{cat}(fsa, fsb) + (1-w_{cat}) * \min_{(i=\{fsa\},\, j=\{fsb\})} Sim_{frame}(fi, fj) \quad (1)$$

Where, Sim_{cat} represents similarity between identified categories of fsa & fsb, and Sim_{frame} represents the frame similarity function between two frames: frame i (fi) and frame j (fj) as mentioned in step 1. Sim_{cat} is defined as inverse of distance between two category nodes in category tree (taxonomy).

Creation of concept clusters is a step towards compressing the video content as it captures distinct macro level independent concepts in video and enable controlled concise representation for them in summary.

Step 4: Concept Ranking

This step ranks the independent concept-clusters created in previous step. Ranking of clusters is performed over its representative category. Weightage for cluster categories are derived from a user model, which maintains a user specific preference score for different levels of categories. Fig. 2 shows concept ranking, where different concepts from video are ranked differently based on user model.

Step 5: Concept Selection

In this step, the top k ranked concept-clusters are selected for formation of summary. 'k' is determined by total number of concept-clusters extracted from video, length of summary desired and number of significant clusters with absolute score above a high_score_threshold (default is 0.9).

significant-clusters = clusters with score > high_score_threshold

*k = min (num_significant-clusters, %s * total_concept_clusters)*

Where, '%s' refers to summary-percentage.

Step 6: Summary Generation

Selected top concept-clusters from previous step are reorganized into summary. This step of summarization process can be highly varying depending upon application needs and user preferences. As a default measure, we adopted following guiding principles:

- Maximum of 3 framesets from each selected cluster are chosen, each contributing 1 (first) frame towards summary.
- Frames from selected clusters are organized in a sequence determined by cluster ranking and their sequence in original video:

(i) Preference should be given to put top ranked clusters ahead in video

(ii) Second preference should be given to retain original sequence of video frames

(iii) Sequence cluster frames according to ordering score as in Eq. 2.

$$cs_{score} = 0.7 * cluster\_ranking + 0.3 * (1 - a / t) \quad (2)$$

Where, cs_{score} represents cluster-sequence-score, 'a' represents original_video_appearance_rank and't' refers to total_frames.

ENTITY IDENTIFICATION USING HIERARCHICAL IMAGE CLASSIFICATION

A hierarchical image classification mechanism has been used to accurately identify the concept associated with a frame (image). The hierarchical classification system attempts to categorize a frame (image) as deep as possible against a designated taxonomy (hierarchy of concept entities). The classification system comprises of N binary image classifiers organized in a tree based hierarchy, where N is no. of nodes in taxonomy and each binary classifier is trained to tell whether image belongs to corresponding category in taxonomy or not.

Each binary classifier is a linear Support Vector Machine (SVM) classifier [9] trained with SIFT dense features from image and quantized using BOF (Bag of Features) model [4], [10]. For entity identification, each image is subjected for classification at the root of classifier hierarchy and classification proceeds further in a top-down fashion. At each step of classification, a decision is made to proceed further down only if classification accuracy at current level exceed a threshold (0.6). The system attempts to classify image as deep as possible in the taxonomy. Fig. 3 represents Top Down Hierarchical Classification.

MANAGING VIDEO SUMMARIES WITH USER'S PREFERENCES

Different users perceive different things from the same video content and hence definition of a video-summary can vary depending on perception & interests of users. In this work, we explicitly try to model different user preferences and include them in a methodical manner to generate dynamic personalized summaries.

A. Representation of User Model

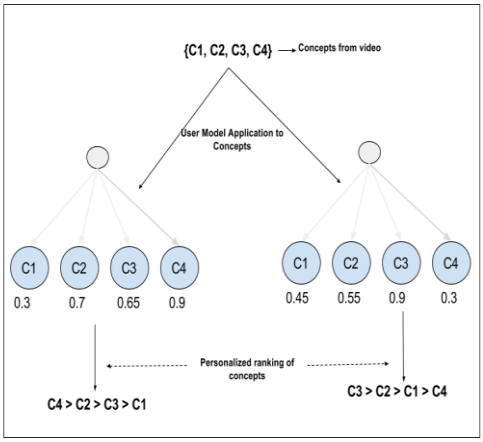
As described in previous section, set of concepts can be represented as a taxonomy - hierarchy of related concept-entities. Users can have varying levels of affinity towards different concepts. Any state-of-the-art generic summarization process in general gives equal weightage to all independent concepts in video. As part of personalized summarization, the current work gives varying user-specific weightage to different concepts resulting in personalized ranking of concepts and accompanying frames.

A user model is a representation of user's preference towards different concepts. For modelling concepts in an organized manner, we define a domain specific taxonomy - hierarchy of concepts. This is a top-down representation of existential relationship between different major concept-entities in a domain (Fig. 4). This taxonomy is crux of our personalized summarization and is utilized in both image-classification and user-model creation. Concretely, a user model for current summarization process is defined as the hierarchy of weight nodes isomorphic to taxonomy, where each user-model node represents score of user's affinity towards corresponding concept node in the taxonomy.

B. Building User Model

There can be many ways to collect user preferences and build such a user-model. Some of the state-of-the-art mechanisms proposed are:

- Explicitly asking users to give their preferences on different topics [11].
- Collectively averaging the preference of video experts and the end users, over a particular video scene or frames [12].
- In addition, following user search & browsing actions can be used to assign & update weights to different nodes of taxonomy
- Frequency of category of search terms (using term classification against same taxonomy).
- Explicit 'like' on videos of different categories.
- Time spent on videos of different categories.



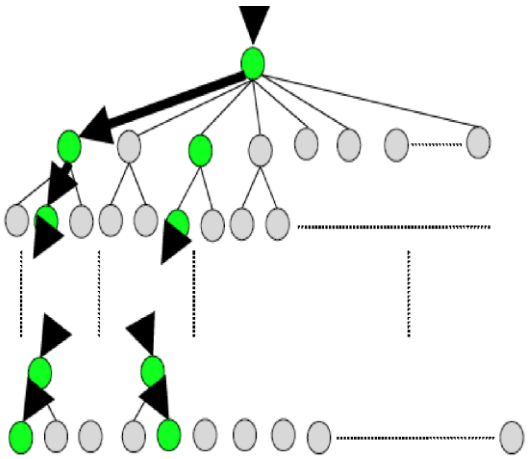


Fig. 2. Concept Ranking **Fig. 3** Top Down Hierarchical Classification

EXPERIMENTS AND EVALUATION

The proposed approach of personalized summarization was evaluated over a video from zoo domain related to animals, birds and reptiles. Concept Taxonomy used in the experiments is depicted in Fig. 4. A set of 3 manually constructed user models interested in birds with different preferences for various types of birds (concepts from taxonomy) were used for evaluation as shown in Table I. Table I shows the preference score on a scale of 0 to 1 for each concept from taxonomy in decreasing order. All other taxonomy concepts not shown are assumed to have zero preference score. Image categorizer system was trained using linear SVM models on SIFT features over image dataset from [13]. Fig. 5a, 5b and 5c represent first frame of personalized summaries for different user model given in Table I with summary length of "5" concepts.

User specific summaries generated by the proposed mechanism were evaluated to measure of how well the presence & ordering (sequencing) of concepts in summary match with user's preferences. For the purpose, concepts present in summary were manually identified along with their associated ordering and compared with expected concept ordering derived from respective user model to generate a customized variance score (V) as defined in Eq. 3:

$$v = \frac{\sum_{\{concepts\_in\_summary\}} \left| R_{cs} - R_{cm} \right|}{\#\, concepts\_in\_summary} \qquad (3)$$

Where, R_{cs} is the order (appearance sequence no.) of concept in summary, R_{cm} is the order of concept as per user model (weighted category tree) order_of_concept_as_per_user_model is ranked order of concept-category in user-model's weighted category tree.

Lower the variance score (V) of a summary, higher is its fitment to the user model. For baseline comparison, the user specific summaries were also compared with generic summary (equal weightage for all concepts).

Results in Table II, III & IV show that proposed mechanism is able to efficiently generate different ordering of concepts in summaries fitting to varying preferences of all 3 user models. Its performance is high particularly when the expected summary length is low, making it a good fit for quickly skimming through multiple videos in a user personalized fashion. Variance scores for generic summaries are not showing any consistent behavior as they attempt to select highly occurring (frequent) concepts in video and may not fit the user expectations in most cases.

Entity
Entity=>Natural
Entity=>Natural=>Sea Beach
Entity=>Natural=>Mountain
Entity=>Natural=>Garden
Entity=>Natural=>Animal
Entity=>Natural=>Animal=>Tiger
Entity=>Natural=>Animal=>Elephant
Entity=>Natural=>Bird
Entity=>Natural=>Bird=>Duck
Entity=>Natural=>Bird=>Ostrich
Entity=>Natural=>Bird=>Owl
Entity=>Natural=>Bird=>Eagle
Entity=>Natural=>Bird=>Turkey
Entity=>Man Made
Entity=>Man Made=>Building_n_House
Entity=>Man Made=>Airplane
Entity=>Man Made=>Airplane=>Fighter Jet
Entity=>Man Made=>Airplane=>Cargo Flight
Entity=>Man Made=>Train

Fig. 4. Experimental Taxonomy

Table I. User Model With Preferences

UserModel	Preferences
A	Duck: 0.95 > Eagle:0.65 > Ostrich: 0.45> Owl:0.35 > Turkey: 0.05
B	Owl:0.95 > Eagle:0.65 > Duck: 0.45 >Ostrich: 0.25> Turkey: 0.05
C	Duck: 0.95 > Ostrich:0.65 > Eagle: 0.45> Owl:0.25 > Turkey: 0.05

(5a)

(5b)

(5c)

Fig. 5. First frame of personalized summaries for user model A (5a), user model B (5b) and user model C (5c)

Table II. Summary Performance For User Model A

Total concepts in Video	**Concepts in Summary**	**Variance Score of personalized summary**	**Variance Score of generic summary**
55	1	0	0
	2	0	1
	3	0	1.66
	5	0	1.6
	10	0.3	0.9

Table III: Summary Performance For User Model B

Total concepts in Video	Concepts in Summary	Variance Score of personalized summary	Variance Score of generic summary
55	1	0	3
	2	0	1.5
	3	0	1.66
	5	0.2	1.2
	10	0.3	0.8

Table IV: Summary Performance For User Model C

Total concepts in Video	Concepts in Summary	Variance Score of personalized summary	Variance Score of generic summary
55	1	0	2
	2	0.5	1
	3	0.33	1.66
	5	0.2	1.4
	10	0.2	0.7

CONCLUSION AND SUMMARY

An approach for personalized summarization using taxonomy based entity identification is presented here. It synergizes identification, compression and user specific ranking of macro concepts present in video for generating effective personalized summaries of desired lengths. Results show that with the availability of defined user model, the proposed mechanism is able to accurately pull user relevant concepts from video in summary with high accuracy, particularity for small summaries. Moreover, the work proposes a methodical way to represent user-model in terms of preference score over concept taxonomy nodes, which helps in its use in effective personalized summarization process and in automated scientific evaluation of personalized summaries as against human perception based evaluation [14]. The work also proposes application of taxonomy based hierarchical image categorization mechanism for explicit concept (entity) identification, which boosts the accuracy of summarization. Apart from its use in summarization task, the proposed mechanism of concept-cluster identification can also find its utility in video indexing & retrieval tasks.

FUTURE DIRECTIONS

The proposed approach relies on image-classification for concept-cluster tagging and hence accuracy of base image classification can result in boosting overall quality of summary. On these lines, we plan to further explore convolution based deep learning solutions for image classification. User model plays a very important role in personalized summarization and hence mechanisms to build and enhance them can play a vital role. We plan to explore different ways in which the models can be created from different user actions including collaborative filtering. We also plan to explore different ways of dynamic summary generation step varying in summary lengths and ordering decided based upon contextual application needs.

REFERENCES

[1] Ejaz, Naveed, Tayyab Bin Tariq, and Sung WookBaik, "Adaptive keyframe extraction for video summarization using an aggregation mechanism." Journal of Visual Communication and Image Representation 23.7, pp. 1031-1040, 2012.

[2] Sujatha, C. and Uma Mudenagudi, "A study on keyframe extraction methods for video summary." In proc. IEEE International Conference on Computational Intelligence and Communication Networks (CICN), 2011.

[3] Joachims, Thorsten, "Text categorization with support vector machines: Learning with many relevant features", Springer Berlin Heidelberg, 1998.

[4] Schmid, Cordelia, "Bag-of-features for category classification," ENS/INRIA Visual Recognition and Machine Learning Summer School Lecture 25-29 July 2011.

[5] Yu, Xiaoyuan, et al., "Automated identification of animal species in camera trap images," EURASIP Journal on Image and Video Processing 1: 1-10, 2013.

[6] https://en.wikipedia.org/wiki/List_of_natural_language_processing_toolkits

[7] Miniakhmetova, Maria, and Mikhail Zymbler, "An approach to personalized video summarization based on user preferences analysis," In proc. IEEE 9th International conference on Application of Information and Communication Technologies (AICT), 2015.

[8] Priyamvada R, Keshaveni, "Frame Clustering Technique towards Single Video Summarization", cognitive computing & information processing (CCIP), 2016 second international conference, IEEE, in press.

[9] http://en.wikipedia.org/wiki/Support_vector_machine

[10] Lowe, David G, "Distinctive image features from scale-invariant keypoints,"International journal of computer vision 60.2: 91-110, 2004.

[11] Ghinea, Gheorghita, et al., "A novel user-centered design for personalized video summarization," In proc. IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2014.

[12] Darabi, Kaveh, and Gheorghita Ghinea, "Personalized video summarization based on group scoring," In proc. IEEE China Summit & International Conference on Signal and Information Processing (China SIP), 2014.

[13] http://vision.caltech.edu/visipedia/CUB-200.html

[14] Kannan, Rajkumar, Gheorghita Ghinea, and Sridhar Swaminathan. "What do you wish to see? A summarization system for movies based on user preferences," Information Processing & Management 51.3: 286-305, 2015.

Authors Profile :

Priyamvada R Sachan: Completed B.E (Instrumentation) from KGCE, Karjat, Mumbai University in the year 2000, M.Tech (VLSI Design & Embedded Sys) from EPCET, Bangalore, Visvesvaraya Technological University in the year 2011 and currently pursuing Ph.D from VTU. Major areas of interest are video processing, machine learning, and image processing.

Keshaveni N: B.E in Electronics & Communication Engineering from Mangalore University in the year 1994. M.Tech in Digital Electronics from B.V.B College of Engineering & Technology Hubli, Affiliated to VTU, Belgaum in the year 2000. PhD (Topic is “Development of Algorithms, Design of Architecture and Implementation of H.264 Advanced Video Encoder”) from Dr. MGR University, Chennai in the year March 2011.