Understanding and Modeling User Interests in Consumer Videos

Ryoma Oami
Multimedia Research Laboratories
NEC Corporation
r-oami@az.jp.nec.com

Ana B. Benitez
Dept. of Electrical Engineering
Columbia University
ana@ee.columbia.edu

Shih-Fu Chang
Dept. of Electrical Engineering
Columbia University
sfchang@ee.columbia.edu

Nevenka Dimitrova
Philips Research
345 Scarborough Road
Briarcliff NY 10510
nevenka.dimitrova@philips.com

Abstract

This paper analyzes the interests of users in viewing and organizing consumer videos. It proposes a taxonomy of relevant concepts with three basic Dimensions Of Interest (DOIs) and effective models to predict the user interest in each dimension. The three DOIs correspond to the objects, the scenes and the events. Our conclusions are backed with an extensive study, in which users were asked to annotate and score the importance of each DOI in short clips of diverse and real consumer videos. Analysis of the user study data reveals high consistency (70%) of the scores across different users, higher importance of objects and events, and independence between objects and events. In addition, we show how heuristic rules and neural networks can accurately predict these scores using camera motion, foreground object and audio information. The automatic and effective prediction of user interests has the potential for improving automatic applications for annotating and summarizing consumer videos, among others.

1 Introduction

In recent years, the increasing popularity of video cameras has stimulated the rapid accumulation of consumer videos. The lack of simple, fast and convenient tools and services to annotate, summarize and manage these consumer video archives, however, has drastically decreased the usability of these videos. Most consumer videos are rarely or never watched after being recorded.

Research on (semi-) automatic summarization and annotation of consumer videos is an emerging field within the multimedia community. The trend in consumer video summarization techniques is to select clips in the video randomly [4] or based on a one-dimensional "importance" score predicted from audio-visual features [1][2]. Probabilistic scene segmentation and clustering based on audio-visual features has also been proposed for accessing consumer videos [5]. The limitation of these approaches is the *a priori* definition of what is "important" or "similar" in consumer videos independent of users. There are several prior works proposing taxonomies and annotation schemes for generic videos [3][6][7][8]; however, none of these approaches have been specially tailored, developed or evaluated for consumer videos with real users. In this paper, we set out to explore what is important in consumer videos from the users' perspective.

This paper proposes a taxonomy of interesting concepts tailored to consumer videos based on a user study. The proposed taxonomy has three basic dimensions of interest (DOI), which correspond to 1) the objects (main characters or entities), 2) the scenes (compositions or aggregations of objects) and 3) the events (actions, changes in objects and scenes, or happenings with special meaning). We conducted an extensive user study to evaluate the proposed taxonomy and, in particular, the three DOIs. Subjects were asked to score the importance of each dimension, and to annotate with free text and/or the taxonomy's concepts several video clips. The video clips were selected from a diverse set of real consumer videos. Analysis of the user study data reveals high consistency (70%) of the scores across different users, higher importance of objects and events, and independence between objects and events.

This paper also analyzes the influence of audio-visual features in DOI scores, and proposes effective prediction models based on simple heuristic rules and neural networks. Our findings point at panning/titling, few large foreground objects or zooming-in, and audio features (music, applause and cheers) to be good indicators of important scenes, objects, and events in consumer videos, respectively. Effective prediction of user interests in consumer videos can greatly advance annotation and summarization tools. For example, if only objects are important in a video clip, the annotation tool can focus on recognizing relevant objects (e.g., people). Consumer video summaries can now be edited following a meaningful and adaptable grammar. A summary can first introduce the main objects and later interleave important events and scenes.

2 Dimensions of Interests

After inspecting several hours of real consumer videos, we realized that people naturally pay attention to multiple aspects while watching consumer videos. In this first analysis, we concluded that objects, scenes and

events were three fundamental Dimensions of Interest (DOI) of users. We define each dimension as follows:

Object: A visible and usually tangible entity (e.g., faces, historical monuments, and pets).

Scene: Aggregation of objects. The scene is the entire appearance or composition of the image rather than the individual objects (e.g., landscape of blooming cherry trees).

Event: Action, change of objects or scenes, or a happening with a particular meaning in the real world (e.g., wedding and baby's first steps).

We then developed a taxonomy of relevant concepts for users in consumer videos having objects, scenes and events as the basic dimensions. Because of space limitations, we do not present the entire taxonomy here. The proposed taxonomy simplifies, adapts and extends several existing taxonomies and annotation schemes [3][6][7][8] to the consumer video domain. For example, the taxonomy used in IBM's annotation tool [3] includes objects, scenes, and events but typical social events in consumer videos (e.g., party) are missing. In addition, we want to emphasize our user-oriented approach for deriving and verifying the salience of concepts in our taxonomy with an actual user study over a non-trivial set of consumer videos.

3 User Study

An extensive user study was conducted to evaluate the proposed taxonomy and the DOIs with real consumer videos and users. This section describes the setup, the methodology and the results of the user study.

3.1 Experimental Setup

Our first task was to obtain a representative data set with diverse and real consumer videos. We collected five hours of home videos taken by three different consumers with ordinary video cameras. These videos showed babies, sightseeing trips, parties, graduation ceremonies, and weddings, among others, covering a wide range of typical consumer video situations.

50 video clips including the important events were selected from the videos. The clips were manually chosen to ensure the user interests remained the same in the entire clip since the user interest can change within a shot (i.e., a segment separated by on-off camera operation). The clips lasted from 5 to 15 seconds.

12 subjects participated in the study. These included five video experts and seven non-video experts. Three of the subjects were the owners of the videos.

3.2 Evaluation Methodology

The subjects were asked to judge and score the importance of objects, scenes, and events in video clips. The scores had three levels: 1, 2, and 3, indicating low, medium and high importance, respectively. The subjects were also encouraged to further annotate the video clips by selecting specific concepts from the taxonomy, or by entering free-text annotations. Each subject was asked to view, score and, optionally, annotate only 15 video clips. Therefore, each video clip was scored and annotated by at least three subjects. This process usually took subjects about 30 minutes.

A brief instruction was given to the subjects before participating in these experiments. The instruction included intuitive definitions of objects, scenes and events; typical examples of video clips and possible scores; and an overview of the annotation taxonomy. The instruction was kept to a minimum because one of the goals of this study was to demonstrate how natural, useful, independent and consistent the taxonomy was.

Table 1. Percentages of consistent video clips per DOI.

	Perfectly consistent	Nearly consistent	Others
Objects	14%	46%	40%
Scenes	38%	30%	32%
Events	38%	42%	20%

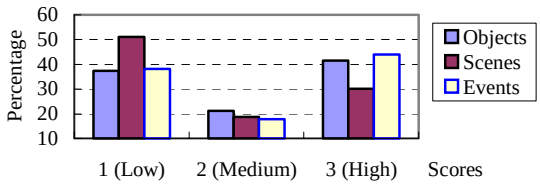

Figure 1. Distribution of score values per DOI.

3.3 Evaluation Results

Table 1 shows the consistency of the scores for objects, scenes and events given to the same video clips by different subjects. "Perfectly consistent" means that the scores were equal for all the subjects; "Nearly consistent" indicates that the scores differ only by one point.

The scores, especially for scenes and events, show high consistency across different subjects (about 70%). The object scores are not so consistent especially because of the different familiarity of subjects with the objects depicted in the video clips. That is to say, some subjects, especially video owners, tended to score known objects (e.g., people) higher than events and scenes, in spite of the existence of multiple foreground objects and actions.

We also analyzed the relationship between the scores assigned to the same video clips independently of the subjects. The correlation between the scores for objects and scenes, scenes and events, and events and objects are -0.654, -0.455, and -0.079, respectively. In other words, the scene scores are mutually exclusive with the object and the event scores, whereas the object and event scores are almost independent. No significant

differences were found between the scores of video expert and non-video expert subjects. The distribution of score values for each DOI is shown in Figure 1. Objects and events were considered more important than scenes in our consumer videos. The scores of the different subjects for each DOI and video clip were averaged for the remaining computations in the paper.

The results of this user study therefore demonstrate the consistency and usefulness of the proposed DOIs as fundamental concepts for consumer videos.

4 Features Affecting the Scores

This section describes the relationships discovered between the scores of the DOIs, and several audio and visual features. Section 5 presents the models built based on these features to predict scores.

The following features were selected through observation for their usefulness in predicting the scores of each dimension:

- Camera motion: zooming in, panning/tilting, object following
- Foreground objects: number of objects and screen area they occupy
- Audio features: music, cheer, applause

These features were manually extracted for the 50 selected clips to assess the feasibility of DOI score prediction using "ideal" features. The camera motion and the audio features are binary: a value of one indicates the existence of the feature. Correlation coefficients between the features and the scores for each video clip and DOI are listed in Table 2. The values were only computed with perfectly and nearly consistent scores.

4.1 Camera Motion

Typical camera motions in consumer video are zooming-in, panning/tilting, and object following. Camera motion that follows objects is distinguished from the simple panning that does not track any specific objects. This is because the purposes of these two camera motions are usually different.

For few foreground objects, zooming-in and object-following camera motion tends to consistently indicate high object importance, according to Table 2. This phenomenon is natural because these camera motions are often used to take a closer look at, or to follow important objects. We have also observed that zooming-in can be used to look at scenes in more detail. This happens when there are a large number of foreground objects. However, in this case, zooming-in is often followed by or simultaneous with, panning/tilting. Panning/tilting camera motion is common when recording scenes rather than objects.

4.2 Foreground Objects

We analyzed the effect of the screen area and the number of foreground objects in the object, scene and event scores. As shown in Table 2, large area (relative to the whole image) of the foreground object(s) is indicative of the presence of important objects and absence of important scenes. This is due to the fact that important objects are likely to appear as closed-ups in the videos. The number of foreground objects has the opposite effect on the scores because the importance of objects decreases with the number of foreground objects.

Table 2. Correlation between features and scores.

Features		Objects	Scenes	Events
Camera motion	Zooming-in	0.256	-0.125	-0.201
	Panning/Tilting	-0.632	0.548	-0.146
	Following	0.366	-0.267	0.142
Foreground objects	Fg object area	0.419	-0.303	0.025
	# of Fg objects	-0.509	0.536	-0.441
Audio	Music	0.076	-0.141	0.293
	Applause	-0.263	-0.284	0.297
	Cheers	0.037	-0.153	0.410

4.3 Audio Features

Table 2 also shows the positive effect of music, applause and cheers in the event scores. This is expected because these features often accompany special and important events such as ceremonies and parties.

5 Constructing Prediction Models

This section describes the models constructed to predict the scores of the DOIs based on the audio-visual presented in section 4. We present two prediction models: a heuristic model and a neural network model.

5.1 Heuristic Prediction Model

A heuristic prediction model was derived from the findings in section 4 with some tedious tweaking. Let the scores for objects, scenes and events be denoted by s_O, s_S, and s_E, respectively. The normalized scores (in the range [0,1]), S_O, S_S, and S_E, are obtained by $S_O = 0.5(s_O - 1)$, and so on.

The proposed heuristic prediction models for the normalized scores are the following:

$$S_O = (1-P)f(n,a) + Z(1-P)\delta(n), \quad (1)$$

$$S_S = P + (1-P)(1-Z)(1-F)\delta(n) + (1-P)(1-\delta(n))(1-f(n,a)), \quad (2)$$

$$S_E = 1-(1-C)(1-M)(1-A), \quad (3)$$

where Z, P, F, C, M, and A denote the zooming-in, panning/tilting, object-following, cheers, music, and applause features, respectively, whereas n and a are the number and area (relative to the whole image) of the foreground objects, respectively. In addition, $\delta(n)$ is a delta function whose value is one when $n = 0$, and zero otherwise; and $f(n,a)$ is a function that indicates the importance of the foreground objects. The shape of the function was derived from the results of the feature analysis and approximated by

$$f(n,a) = g_1(n)g_2(a), \quad (4)$$

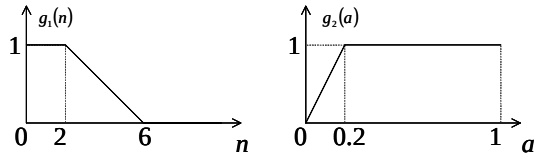

Figure 2. Functions used in the prediction model.

where $g_1(n)$ and $g_2(a)$ are shown in Figure 2.

Regarding the object score, it increases with the area of the foreground objects (first term in (1)) or the presence of zooming-in (second term), in either case without panning/tilting. On the other hand, the object's importance decreases with the number of foreground objects (first term), as explained in section 4. The presence of panning/tilting is the clearest indicator of high scene scores (first term in (2)). The second term indicates that the scene becomes important if there are no foreground objects or camera motion. The third term is almost opposite to the first term in (1) without panning/tilting camera motion: the scene score is high when there are many foreground objects but their area is small. The event importance increases considerably with the presence of music, applause, and cheers (3). The event score is thus the logical sum of these features.

The Mean Square Error (MSE) of the predicted values using the models above are 0.205, 0.475, and 0.832 for objects, scenes, and events, respectively, using the consistent video clips (0.366, 0.515 and 0.863 using all the video clips). The prediction error for the event score is considerable but the models can predict object and scene scores satisfactorily.

5.2 Neural Network Prediction Model

We also built a prediction model based on a neural network. First, two video experts generated the object, scene and event scores for 200 additional video clips because the initial 50 video clips were considered insufficient to train and test the neural network. We built a neural network that had 8 inputs (one per feature in Table 2), three outputs (one per object, scene, and event scores), and two hidden layers with units shared by the three outputs. A 10-fold cross validation was conducted to determine the optimal number of hidden units. The MSE in predicting the object, scene, and event scores is 0.365, 0.446, and 0.558, respectively.

5.3 Discussion

The two prediction models can be compared in terms of prediction accuracy and model understanding. Whereas the neural network can predict the event score more accurately, the heuristic model's predictions for object scores have less error. The performance of both models for predicting the scene score is comparable. In terms of model understanding, a clear advantage of the heuristic model is that the relationships between the features and the scores are explicit and known. This makes it easy to adjust the model in accordance with user preferences or video characteristics.

6 Conclusions

This paper has proposed a taxonomy of relevant concepts with three basic dimensions of interest (objects, scenes and events) tailored to consumer video based on a user study. The user study has shown high consistency and usefulness of object, scene and event demonstrating their suitability as basic dimensions of user interests in consumer videos. In addition, further analysis of the user study data has revealed high consistency (70%) of the scores across different users, higher importance of objects and events, and independence between objects and events. Models that can accurately predict these scores have also been proposed based on simple heuristic rules and neural networks. Our findings point at panning/titling, few large foreground objects or zooming-in, and audio features to be good indicators of important scenes, objects, and events in consumer videos, respectively.

Effective prediction of user interests in consumer videos can greatly advance annotation and summarization tools. For example, if only objects are important in a video clip, the annotation tool can focus on recognizing relevant objects (e.g., people). Consumer video summaries can also be edited following a meaningful grammar adaptable to user preferences. A summary can first introduce the main objects and later interleave important events and scenes.

7 References

[1] J. R. Kender and B. L. Yeo, "On the Structure and Analysis of Home Videos", ACCV, 2000.

[2] Y. F. Ma, L. Lu, H. J. Zhang, and M. Li, "A User Attention Model for Video Summarization", ACM Multimedia, 2002.

[3] B. Adams et al. "IBM Research TREC-2002 Video Retrieval System", 2002.

[4] Rainer Lienhart, "Abstracting Home Video Automatically", ACM Multimedia, 1999.

[5] D. Gatica-Perez, A. Loui, and M.T. Sun, "Finding Structure in Home Videos by Probabilistic Hierarchical Clustering," IEEE Trans. CSVT, 13, 6, pp. 539-548, 2003.

[6] P. Salembier, and J. R. Smith, "MPEG-7 Multimedia Description Schemes", IEEE Trans. CSVT, 11, 6, pp. 748-759, 2001.

[7] C. Lindley, "A Video Annotation Methodology for Interactive Video Sequence Generation", Conf. on Digital Content Creation, 2000.

[8] M. Davis, "Media Streams: Representing Video for Retrieval and Repurposing", Ph.D. Thesis, MIT, 1995.