In Proc. IEEE International Conference on Image Processing, ICIP 2002, Sept. 22-25, Rochester, NY, USA

LEARNING PERSONALIZED VIDEO HIGHLIGHTS FROM DETAILED MPEG-7 METADATA

Alejandro Jaimes [*], Tomio Echigo [Ψ], Masayoshi Teraguchi [Ψ], and Fumiko Satoh [Ψ]

* Electrical Engineering Department
Columbia University
New York, NY 10027, USA

[Ψ] Multimedia Group
IBM Tokyo Research Laboratory
Yamato, Kanagawa 242-8502, JAPAN

ABSTRACT

We present a new framework for generating personalized video digests from detailed event metadata. In the new approach high level semantic features (e.g., No. of offensive events) are extracted from an existing metadata signal using time windows (e.g., features within 16 sec. intervals). Personalized video digests are generated using a supervised learning algorithm which takes as input examples of important/ unimportant events. Window-based features are extracted from the metadata and used to train the system and build a classifier that, given metadata for a new video, classifies segments into important and unimportant, according to a specific user, to generate personalized video digests. Our experimental results using soccer video suggest that extracting high level semantic information from existing metadata can be used effectively (80% precision and 85% recall using cross validation) in generating personalized video digests.

1. INTRODUCTION

Multimedia users can be characterized by at least two important factors. The first one relates to their *limited resources* in terms of time and the devices they have access to when using multimedia data. There are many possibilities when choosing a viewing device, for example, ranging from computers connected to high bandwith networks, to handheld devices with limited computational and bandwidth resources. The second aspect relates to *personal preferences* in terms of the information that each user wants to view and the way in which it is used.

The new MPEG-7 standard [4], which aims at standardizing tools for describing different aspects of multimedia at different levels of abstraction, facilitates management of multimedia data and is likely to contribute to the explosion of the availability of metadata. Given the importance of semantic information, a large part of current and future MPEG-7 descriptors are likely to be semantic (i.e. textual).

Given these issues (limited resources, personal preferences, and availability of semantic metadata), it is extremely important to develop approaches that efficiently and effectively *summarize* multimedia data in a *personalized* way making use of high level information available in the metadata.

IBM T.R.L. has been working on a project to address these issues. In particular, a system has been constructed to generate personalized video digests so that such digests can be viewed on handheld devices (e.g., PDAs, cellular phones). The system consists of an MPEG-7 annotation tool and a video digest generation module. The annotation tool allows an operator to add detailed textual metadata to post-production videos. The video digest generation module uses individual user profiles to select segments of a video that are relevant to the user's interests (Figure 1). Unlike video summaries, digests are meant to include only the highlights of a video (not summarize its entire contents).

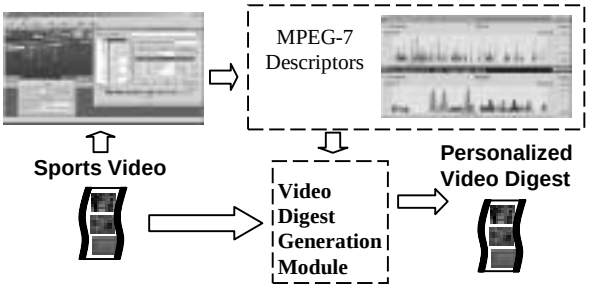

Figure 1. System overview.

In this paper, we present a new approach for generating personalized video digests from detailed event meta-data. In particular, we extract high level semantic features from the metadata and use supervised learning to generate personalized digests.

We treat metadata as a discrete time signal. An *activity window* is determined by a time interval [*t1, t2*]. Features are extracted (e.g., number of offensive events, etc.) that semantically represent the importance of the events in the time window. A user labels highlights in a set of training videos, generating a set of relevant and non-relevant segments. A supervised learning algorithm is then used with the training set to construct a classifier that, given new metadata for a video (soccer), can classify segments (windows) into relevant and irrelevant according to the specific user's interests. Only relevant game segments are included in the personalized digest.

Research on video summarization has been extensive. In many approaches, the goal is to create an efficient overview of the *entire* contents of a video (e.g., grouping similar video shots [6]). Our work differs from these in several aspects: (1) we focus on selecting only the *most relevant* sections of content; (2) we seek generation of individual (rather than general) summaries that depend on *personal preferences*; (3) our framework makes strong use of *detailed* event metadata. It also differs from approaches that do not use learning or do not produce personalized digests (e.g., [1]).

The work in this paper also differs from our previous work [2][5]: (1) we automatically extract high level semantic

information from the metadata; (2) we replace manually constructed user profiles with a supervised learning algorithm.

In I.R. [3], the stages of information extraction, term weighting, and relevance ranking are related to our scenario. Unlike in I.R., however, there is no specific query, and we deal with very small amounts of data (not millions of samples). Furthermore, differences in data distribution are likely to invalidate some of the assumptions made in many I.R. approaches.

2. EVENT METADATA

For a video, metadata is manually input by an operator using an MPEG-7 annotation system [5]. The system, which contains a list of domain-specific event names (e.g., goal, corner-kick, etc.), allows annotation of post-production videos in almost real time. In soccer videos, each annotation consists of a time stamp, and an event name (Table 1).

Event Name	Freq	O/D	Event	Freq	O/D
Red card	1	-	Goal kick	64	D
Replay corner kick	3	-	Replay foul	75	-
Replay free kick	3	-	Replay shoot	100	-
Bar/Post	5	O	Keeper cut	155	D
Missing pass	7	-	Shoot	165	O
Replay injury	7	-	Free kick	178	O
Replay keeper cut	7	-	Throw in	195	-
Goal	11	O	Foul	205	-
Replay centering	19	-	Ball cut	250	D
Yellow card	22	-	Through pass	304	O
Offside	24	O	Centering	323	O
Replay goal	32	-	Long pass	337	O
Player change	35	-	Pass cut	792	D
Shoot block	35	D	Dribble	1250	-
Clear	57	D	Pass	1687	-
Corner kick	57	O			

Table 1. Some events, offensive/defensive categories, and frequencies for all the games used in section 4.

Manual annotation of content is expensive, but in many cases cannot be replaced by automatic techniques[1]. This is particularly true when the annotations are *semantic* (e.g., events) as opposed to *syntactic* (e.g., color). Automatic scene cut detection, for example, has been applied successfully in several domains. Accurate automatic detection of *semantic* events in videos, however, remains an open research area with many challenging issues.

In manual annotation there is a tradeoff between the detail of the annotations and the cost (in terms of human effort). The goal, then, is to minimize the work performed by the operator annotating the videos, while maximizing the utility of the annotations generated. The annotations themselves, therefore, must be as simple as possible, but provide enough information to be useful. Therefore, we propose extracting domain specific high-level semantic information from multimedia metadata. This increases the usefulness of the metadata without increasing the indexing cost, and takes advantage of important domain specific knowledge.

In many sports, for example, it is possible to group events into several categories (see some examples in Table 1).

- *Start action:* ball not in play before the event, but in play immediately after the event.
- *End action:* ball is in play before the event, not in play immediately after the event.
- *Ball not in play:* ball not in play before the event, not in play after the event
- *Ball in play:* ball is in play before the event, and maybe after the event
- *Determine location:* the event name carries some information about where it occurs.
- *Affect outcome:* the event creates a permanent change in the game.
- *Change score:* the event changes the score of the game.
- *Offensive:* the event indicates an offensive action.
- *Defensive:* the event indicates a defensive action.

Grouping events into categories can be very useful for generating personalized video digests (and for queries) because it captures important high level information that may be easily understood by users (e.g., an “active” segment is a segment in which the ball is in play most of the time). For example, a user could request a video digest that includes "all offensive events." Furthermore, the approach can be easily generalized to other domains (e.g., the same categories can be used in other sports; similar categories can be used in other types of videos).

3. VIDEO DIGEST GENERATION

Once the metadata has been created, it is used to generate personalized video digests. In previous work [2][5], a user profile (a list of event-importance weight pairs: e.g., goal 10, corner-kick 7, pass 3, etc.) was used to rank events according to importance. The profile (specific to a user or group of users), was then used to score individual events within a specific video. Events in the video whose importance score was higher than a given threshold (and fit a given time budget) would be included in the video digest.

The construction of profiles is difficult, however. Often the difference in importance between events is difficult to determine (e.g., corner kick vs. free kick). In the new approach presented in this paper, content profiles are not needed because user preferences are learned automatically. In addition, new high-level semantic features are extracted from the existing metadata. These features, described next, are based on the categories of the previous section.

1. Feature Extraction

A user profile is a list of event-importance pairs: up = $\{ge_1, ge_2, ge_3, \ldots, ge_m\}$ where $ge_q = \{label_q, weight_q\}$. A video S, can therefore be described as follows:

[1] In our framework, the content is manually annotated, but the digests are generated automatically.

- S = {{label$_1$, time-stamp$_1$, importance-weight}, {label$_2$, time-stamp$_2$, importance-weight$_2$},…,{label$_n$, time-stamp$_n$, importance-score$_n$}}

Each sample in the video sequence corresponds to an event and has a label, a time stamp and an importance score based on the user profile. A discrete-time system can be defined which maps the metadata input sequence m[*n*] to an output sequence o[*n*]: o[*n*] = T {m[*n*]}. In our approach, a window is defined over a time interval [t_1, t_2] (where $t_2 > t_1$). For example (Figure 2) W_i = {{goal, 13:12:05, 10}, {corner-kick, 13:11:50, 9},….,{goal, 44:05:04, 10}}

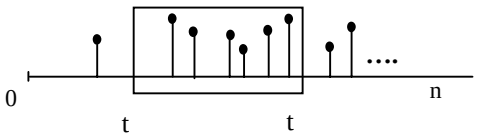

Figure 2. Metadata signal.

Previous work [2][5] was based solely on user profiles and thresholds were used to generate the digests. Instead, in this paper, we propose using features, extracted within the window, that represent important domain specific semantic information that is useful for generating digests[2].

- No. of events inside the window: m
- No. of location events: $\sum_{i=0}^{i<m} Le_i$ where Le_i = 1 if e_i is a location event and Le_i=0 otherwise.
- No. of interruptions: $\sum_{i=0}^{i<m} Ie_i$ where Ie_i = 1 if e_i is an event that stops action and Ie_i=0 otherwise.
- No. of defensive events: $\sum_{i=0}^{i<m} De_i$ where De_i = 1 if e_i is a defensive event and De_i=0 otherwise.
- No. of offensive events: $\sum_{i=0}^{i<m} Oe_i$ where Oe_i = 1 if e_i is an offensive event and Oe_i=0 otherwise.
- Play time: $\sum_{i=0}^{i<m} t(e_i)$ where $t(e_i)$ is a function that returns the length of time for which the ball is in play for the event e_i. This value can be computed given the classification of events into those that start action, end action, etc. If the event is a start action event, for example, the time in play for the event corresponds to the time between that event and the next event e_{i+1}.

The metadata, then, is treated as a one dimensional signal for which window-based high level semantic features are extracted. The window is shifted over the original signal (one point at a time) to produce new values for each sample. The resulting signal is used to generate the video digests.

[2] Note that in this framework, visual features within each window could be easily incorporated.

2. Learning User preferences

The system consists of a training phase and a classification phase (Figure 3). During training, a user watches a set of training videos and labels the highlight events he is interested in. It is only necessary for the user to tag interesting events with a "yes" label. Unmarked events are used as negative examples.

Features[2] (f_1,..,f_n within a window of length t) are extracted from the corresponding metadata signals, generating a training set. In the training set, positive window examples are those that contain events chosen by the user as highlights. The remaining windows constitute the negative examples. The training set is used by a supervised learning algorithm to learn a binary classifier that can determine highlight preferences for a specific user (or group of users if the training comes from several users- e.g., fans of the Japanese national team).

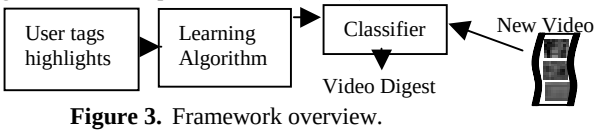

Figure 3. Framework overview.

During classification, metadata from a new video is input into the system. The same features (using a window of the same size) that were extracted in the training phase are extracted from the new metadata signal, and the classifier learned by the supervised learning algorithm is applied. The output of the classifier is a set of segments, each with either a positive (highlight) or negative label (not highlight). Only highlight events are included in the personalized video digest.

There are several advantages of this approach over previous work [5][2]. First, user profiles (event-importance pairs) are not necessary. Second, existing metadata is used more efficiently because higher level features are extracted. Third, preferences are learned automatically. In addition, the high level features proposed could be used to browse or query existing videos (e.g., "active segments", etc.).

4. EXPERIMENTAL RESULTS

Six post-production soccer games (Table 2) were collected and annotated by one of the authors using the MPEG-7 annotation system [5]. One of the other authors watched each of the videos and manually selected his personal highlights.

All of the features of the previous section were extracted from the metadata of each of the games using overlapping windows. The database was split into several training sets (Table 3).

GAME	**TIME (mins)**	**No. events**	**Avg. / min.**
FUKUOKAvs.ICHIHARA	101	1,800	18
JAPAN vs. JAMAICA	97	2,041	21
JAPAN vs. PARAGUAY	101	1,798	18
JAPANvs.YUGOSLAVIA	100	1,820	18
JAPAN vs. CROATIA	93	1,581	17
ITALYvs.FRANCE	96	1,890	19
TOTAL	588	10,930	18

Table 2. Data summary.

Set	Description	Training Pos.	Training Neg.	Samples
A	All highlights	50%	50%	3156
B	67% of the data	50%	50%	1986
C	All examples	14%	86%	10917

Table 3. Training set descriptions.

First, we performed an experiment to determine how well a human, analyzing only the metadata, would do in selecting the highlights of the videos. This was done on the entire database (set C). A simple threshold was used to determine the human subject's performance. Then, we used various supervised learning algorithms from [8] on the different training sets using varied window sizes (e.g., 8, 16, 32 seconds). The choice of window size depends largely on the viewing device and domain- for soccer video on PDAs and cellular phones, for example, we found 16 seconds to be suitable for generating short digests with enough detail.

Results from these experiments are very promising. The human performance on selecting highlights by viewing only the metadata on the entire set yielded 89% precision and 16% recall. Using 16 second windows and cross-validation on set B produced 80% precision and 85% recall. Although it is difficult to directly compare human performance with our approach, such a measure highlights the difficulty of selecting highlights from the given metadata.

Set	**A**	**A**	**B**	**B**	**C**	**C**
Method	**Prec.**	**Recall**	**Prec.**	**Recall**	**Prec.**	**Recall**
1-Nearest N.	78%	69%	72%	80%	54%	36%
3-Nearest N.	80%	77%	80%	80%	60%	35%
Neural N.	79%	83%	**80%**	**85%**	63%	40%
Human	-	-	-	-	89%	16%

Table 4. 10-fold cross-validation performance results on training sets using 16 second windows and various learning algorithms.

Using 67% (4 games) of the set for training and testing on the remaining 33% (2 games) we obtained 44% precision and 75% recall using a Neural Network classifier (Table 5). These results compared to those obtained using cross-validation show the benefit of having a larger training set (90% of the data in the case of cross-validation; equally distributed positive and negative training examples).

Training Set	**B**	**B**
Algorithm	**Precision**	**Recall**
1-Nearest Neighbor	19%	87%
5-Nearest Neighbor	42%	74%
NN	44%	75%

Table 5. Independent test set results.

The extraction of high level semantic features from the metadata is beneficial not only in terms of performance, but in that it eliminates the need to use content profiles to score individual events. In addition, the extracted features can be used to perform other types of queries.

Using a supervised learning algorithm also improves performance. The specific choice of algorithm, however, depends on the application- in the case of mobile devices (PDAs, phones), a Neural Network might be more suitable because of its faster classification speed (compared to a nearest neighbor approach).

5. CONCLUSIONS AND FUTURE WORK

A new way of generating personalized video digests was presented. The approach is based on extracting new high level domain specific features from detailed event metadata, and on using those features in a supervised learning algorithm. A user selects his personal event highlights in a set of videos. Features are extracted from the training set to yield a classifier that determines, from detailed event metadata of a new game, the highlights according to the user's preferences. Our experiments suggest promising results. Using cross-validation on a set of 6 soccer videos, 80% precision and 85% recall was achieved for a specific user's personal preferences. Results were lower using an independent test with a smaller training set (44% precision and 75% recall). An experiment in which a human subject chose highlights viewing only the metadata signals, yielded 89% precision and 16% recall, highlighting the difficulty of selecting highlights from detailed event metadata.

Future work includes using visual features in addition to metadata features, and applying the methodology in other domains.

6. REFERENCES

[1] N. Babaguchi, "Towards Abstracting Sports Video Highlights," Proc., *IEEE Intl. Conference on Multimedia and Expo 2001*, Vol. 3 pp. 1519-1522, Tokyo, Japan, 2001.

[2] T. Echigo, K. Masumitsu, M. Teraguchi, M. Etoh, and S. Sekiguchi, "Personalized Delivery of Digest Video Managed on MPEG-7," Proc., *IEEE International Conference on Information Technology: Coding and Computing*, pp. 216-220, April 2001.

[3] W.B. Frakes and R. Baeza-Yates, eds. *Information Retrieval: Data Structures and Algorithms.* Englewood Cliffs, NJ, Prentice Hall, 1992.

[4] MPEG Requirements Group, "MPEG-7: Context, Objectives and Technical Roadmap, V.12", *ISO/IEC JTC1/SC29/WG11 MPEG99/N2861*, Vanc., July 1999.

[5] K. Masumitsu and T. Echigo, "Meta-Data Framework for Constructing Individualized Video Digest", Proc., *IEEE Intl. Conf. On Image Processing 2001*, Vol. 2, pp. 390-393, Thessaloniki, Greece, 2001.

[6] M. Yeung, B.L. Yeo, et al., "Extracting Story Units from Long Programs for Video Browsing and Navigation," Proc., *IEEE Third Int. Conf. On Multimedia Computing and Systems*, pp. 296-305 1996.

[7] S. Uchihashi et al. "Video Manga: Generating Semantically Meaningful Video Summaries," Proc. *ACM Multimedia 99*, pp. 383-292, Orlando, FL, Nov. 1999.

[8] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufman, New York, 1999.