Beyond Audio and Video Retrieval: Towards Multimedia Summarization

Anonymous ICMR 2012
Submission

ABSTRACT

Given the deluge of multimedia content that is becoming available over the Internet, it is increasingly important to be able to effectively examine and organize these large stores of information in ways that go beyond browsing or collaborative filtering. In this paper we review previous work on audio and video processing, and define the task of Topic-Oriented Multimedia Summarization (TOMS) using Natural Language Generation: given a set of automatically extracted features from a video (such as visual concepts, ASR transcripts, etc.) a TOMS system will automatically generate a paragraph of natural language ("a recounting"), which summarizes the important information in a video belonging to a certain topic area, and provides explanations for why a video was matched, retrieved, etc. We see this as a first step towards systems that will be able todiscriminate visually similar, but semantically different videos, compare two videos and provide textual output or summarize a large number of videos at once. In this paper, we introduce our approach of solving the TOMS problem. We extract Visual Concept features and ASR transcription features from a given video, and develop a Template-Based Natural Language Generation (NLG) System to produce a textual recounting based on the extracted features. We also propose possible experimental designs for continuously evaluating and improving TOMS systems, and present results of a pilot evaluation of our initial system.

General Terms

Management, Measurement, Documentation, Performance, Design, Experimentation, Human Factors, Languages.

Keywords

Multimedia Summarization, Event Detection and Recounting, Natural Language Generation.

1. INTRODUCTION

Consumer-grade video is becoming abundant on the Internet, and it is now easier than ever to download multimedia material of any kind and quality. With cell-phones now featuring video recording capability along with broadband connectivity, multimedia material can be recorded and distributed across the world just as easily as text could just a couple of years ago. The easy availability of vast amounts of text gave a huge boost to the Natural Language Processing (NLP) research community, which was critical in order to organize the amount of information that was suddenly available. The above-mentioned multimedia material is set to do the same for multi-modal audio and video analysis and generation, and we argue that NLP can play a big role in organizing this information, and making it accessible.

State-of-the-art techniques for accessing audio and video material are mainly designed to facilitate browsing of a video, and generate recommendations based on collaborative filtering. In our vision, a good textual summary will help the user obtain maximal information from the video, without having to watch the video from beginning to the end. When placed in mouse-over "tooltips", or similar context-sensitive elements of a graphical user interface, text can enhance the browsing process. A single summary could for example describe a whole set of similar videos, or a summary could describe why a specific video is different from other, related videos. This will be particularly useful to quickly spot "false positives" in retrieval applications at a semantic level, rather than a low-level feature level. Finally, a summary could compare two videos, and explain how these videos are different. When broadcast on Twitter (which is text oriented for efficiency), RSS feeds, or placed on banners, a good text summary could elicit interest, which will then lead to a browsing session. In addition to facilitating browsing and analyzing of a video, another important goal of our proposed summarization approach is to help the user understand why, with respect to some external information, a video was classified into a certain category, or why the video was retrieved in response to a certain query. In the process, information from audio and video modalities will be fused and presented in a unified way. Rather than only returning audio/video content matching the query, the advantage of Topic-Oriented Multimedia Summarization (TOMS) lies in its ability to merge evidence from various modalities, including visual semantic concepts, Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), and semantic audio concepts, to present the systems' results in a natural, intuitive format.

In order to develop TOMS, we start by building a system capable of generating a passage of human readable text to describe (recount) the objects, people, and activities that can be observed in a video. We do this on a dataset of videos with topic labels [12] (both manually assigned and automatically derived labels are available), so that the recounting is geared towards discussing the evidence and reasoning with respect to the event(s) that defines the topic. These events are defined in a so-called "event kit", which also contains a textual description of important objects and actions that make up each topic.

ICMR'12, Jun. 5–8, 2012, Hong Kong.

Given a series of features extracted from the video, the first step of the TOMS system is to select the features that contain the most salient information about the video, and its topic. For example, if a video is showing a wedding ceremony, a good recounting system should extract the important features from the video and generate a passage that talks about the objects, the people and the sound (speech and music) in the video. A brief example would be:

The video is about a wedding ceremony. We saw a crowd, people in black and white, and flowers in the video. We also heard wedding music and people talking about the beautiful bride.

At this early stage, we incorporate the two most salient types of features, out of four, which our group currently uses for video classification: (1) video semantic concepts ("vehicle", "animal", "body parts", etc.) represent visual information, while (2) English speech recognition output captures the information in the audio signal. We are looking at algorithmic approaches to unify and convert these feature labels to text, which is suitable for downstream processing. We plan to also use (3) OCR features, and (4) audio semantic concepts ("engine", "noise", "music"). Eventually "actions" and other semantic temporal and spatial complex constructs will be included in the summaries. It should be emphasized that we are not attempting to extract specific identities of people or objects in this work.

After feature extraction and selection, a Template-based Natural Language Generation (NLG) system produces recounting passages for the video. A topic-independent planner module creates the summary by suitably concatenating the output of multiple, specialized NLG modules, which are each using a unique, manually written natural language template, to generate a sentence about observed evidence, and its importance. While the NLG module is entirely rule-based for now, we envision a system in which mappings from multimedia evidence to natural language summaries are automatically learned from data, e.g. by adapting techniques developed for Machine Translation (MT), such as used in the MOUNTAIN system [9]. This paper describes our initial system, as well as evaluation strategies that can be used to empirically evaluate and improve future systems.

The paper is organized as follows. Section 2 reviews related work in the area of audio and video summarization. In Section 3, we briefly present the multimedia features that we extract, and use for classification. Section 4 describes the NLG system. The user study based experimental design and results are described in Section 5. Finally, Section 6 concludes the paper, summarizing findings and outlining future work.

2. RELATED WORK

2.1 Audio Summarization

Automatic audio summarization is an ongoing research pursuit, which relies either on algorithms to identify and remove redundancy, for example in music or noise, or first turn speech into text, and then employ text summarization methods. The peculiarities and potential ambiguities of decoded audio such as high recognition error rates, lack of syntactic boundaries, etc., need to be addressed specifically for extracting summary information from audio for content-based browsing and skimming. Valenza et al. [17] were one of the first to present a method combining acoustic confidence measures with information retrieval and extraction techniques, in order to obtain accurate and readable summaries of broadcast news programs. They also demonstrated how extracted summaries, full-text speech recognizer output and audio files can be linked together usefully with a graphical user interface.

Generally, speech summarization can be performed by simply extracting salient words [7] or sentences from the original data, or by synthesizing new representations from the original data. The second case is of course more difficult, harder to evaluate, but also potentially more useful, because the information representation cannot only be compact, but also targeted, clean, and easy to understand. It is our goal for multimedia summarization to achieve similar progress with respect to video retrieval.

Other relevant work investigates how "noteworthy utterances" can be extracted from meetings, and how speech summarization is possible based on non-textual features alone. In all works, evaluation (i.e. how much information is retained at a given compression ratio, and how easy is it to comprehend) has played a major role in the development, with the consideration and fusion of multiple information sources proving helpful [10][13].

2.2 Video Summarization

The large amount of multimedia data available on the Internet is making video content summarization methods increasingly important. Truong et al.'s article [14] about video abstraction describes the techniques targeting video data from various domains (e.g. online videos, movies, critiques, documentaries, news, home recordings, etc.) that were developed to summarize information from the video and to present to the user as surrogates. Some services use a single keyframe to represent the video (like Yahoo and Alta Vista), while some provide a context-sensitive keyframe list of the video (like Google).

Christel, et al. [3][4] in the Informedia group at Carnegie Mellon University have conducted research in user interface designs for video browsing and summarization. In their experiments, Christel et al. use single thumbnails, thumbnail storyboards, playable video skims, and complex "video collages" featuring multiple synchronized information perspectives as summarization tools. They report in their experimental results the merits of discount usability techniques for iterative improvement and evaluation. They also discuss the structure of formal empirical investigations with end users that have ecological validity while addressing the human computer interaction metrics of efficiency, effectiveness, and satisfaction. These summaries were evaluated as either informative summaries providing succinct descriptions of the original videos, or as indicative summaries for judging relevance given a particular search query. In shot-based retrieval experiments, visually dense storyboard presentations worked best, but recounting for justifying event-based retrieval was not investigated.

Previously, summarization of a video typically meant a graphical representation such as visually rich (context sensitive) storyboards, were being used to help the browsing process (as with Open-Video Archive [11]). Video summaries had a temporal aspect in terms of playable audio-visual material [6][11]. The informative summary for a video exploiting both audio and video information was improved with a maximal marginal relevance algorithm working across video genres [10]. In this work, our focus is on static presentation appreciated all at once, emphasizing textual summaries as done by Ushiku et al. [16]. In the evaluation we present later in this paper, we will investigate an indicative approach to summarization (providing evidence for membership in a topical class), as well as an informative summarization.

Recently, Tan et al. [15] have proposed utilizing audio- visual concept classifiers obtained based on the recognition techniques to generate textual descriptions of video contents. In their approach, 2D static SIFT, 3D spatial-temporal interest points (STIPs) and

MFCC audio descriptors have been used to extract audio-visual concept features from the videos. Then a rule-based approach generated textual descriptions after manually defining a template for each concept. For the evaluation, they conducted a user study by asking 43 human evaluators to rate each text description on a one (negative) to five (positive) scale. One-third of the ratings were three to four, while half of them were five. Since the evaluation was completely subjective, the informative conclusion of the result was limited (efficacy and efficiency were not addressed). Also, ASR and OCR features were not applied in this work. The template approach, which is directly linked to complete events, appears to not scale well to large amounts of video. Our work attempts to address some of these limitations.

3. MULTIMEDIA FEATURE DEFINITION AND EXTRACTION

In our task, each video is labeled as one of 10 events (topics) from the TRECVID 2011 Multimedia Event Detection (MED) task and database [12]:

1. Birthday_party,
2. Changing_a_vehicle_tire,
3. Flash_mob_gathering,
4. Getting_a_vehicle_unstuck,
5. Grooming_an_animal,
6. Parade,
7. Making_a_sandwich,
8. Parkour,
9. Repairing_an_appliance,
10. Working_on_a_sewing_project

Currently, about 1000 videos have manual (reference) topic labels, while several 10000 videos are available, and have automatically generated topic labels from the MED systems. TOMS is designed to provide first an indicative summary that provides evidence for membership in one of the 10 MED events; and second to generate a recounting summary passage to present the features and concepts that have been detected in the video. A TOMS system can therefore create text for reference topic labels, but also for automatically generated topic labels, which might well be wrong.

In this work, we use the most salient Semantic Indexing (SIN) and ASR features, although it is straightforward to extend the system.

3.1 Video-Level SIN Feature

We employ the visual concept detector to index all extracted keyframes from the given video. For each keyframe we calculate scores for each of the 346 visual concepts. Note that these visual concept detectors are SVM classifiers trained over the SIN task in TRECVID 2011 using MOSIFT and CSIFT features to describe keyframes [1]. To determine the video-level semantic indexing, we simply take the average of the keyframe-level SIN for all keyframes within the video. Note that we evaluate different ways to determine the video-level SIN representation in the context of the MED task such as taking the max, median, mean, and etc. of the keyframe-level SIN and the experimental result shows the superiority of the taking the average method to merge the keyframe-level SIN and generate the effective video-level representation.

3.2 Ranking Visual Concepts

As an example for our approach to compute as many aspects of the re-counting automatically, rather than manually coding it in (ad-hoc) rules, we present the way in which we extract the list of features to mention in a recounting, using a bipartite graph:

For each video, we aim to rank the detected visual concepts, in order to mention the most important ones in the recounting. If we rank the videos according to their determined probabilities, in some cases general concepts such as “human”, “indoor” and etc., which generally have higher probabilities than others, are placed in high-ranked positions, while they might not be discriminative and informative for the event of interest. As the trained concept detectors generally have low precision (mean precision of around 0.17) there are cases in which computed probabilities are not basically reasonable. Therefore sorting them according to their probabilities can end up with noisy and meaningless ranked lists.

To cope with this problem and provide a more accurate visual concept-ranking list, we take both discrimination and relatedness of visual concepts into account through two steps. Considering the machine capability to detect different visual concepts, first we remove less discriminative visual concepts. Second, we take the human perception into account and re-rank the remaining visual concepts with respect to the manually determined ground truth for each event. We briefly explain each step in the following.

3.2.1 Visual Concept Discrimination Analysis

First we determine the global rank list of visual concepts considering their distinguishing power based on which less discriminative concepts can be removed. To do so, we explore pair-wise relationship between training videos and concepts using graph propagation methods [2] to rank visual concepts for each event in the descending order of their discriminative power.

As shown in Figure 1, $G=(V,C,E,W)$ is bipartite graph between training videos and concepts, where V is the node set for training videos, C is the node set for concept, E is the edge set and the edge is weighted by W_{ij}. W_{ij} is the concept c_j's prediction score on video v_i.

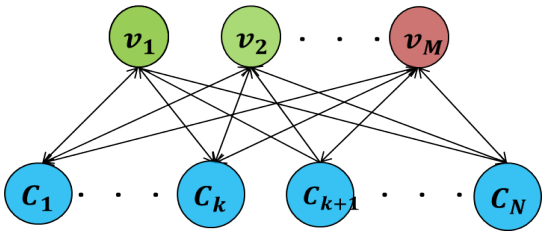

Figure 1. Bipartite Graph Construction for determination of relevant features to mention in recounting.

The propagation process in the bipartite graph can be written as:

$$\begin{cases} f_{t+1}^{c} = \alpha \tilde{W}^{T} f_{t}^{v} + (1-\alpha) y^{c} \\ f_{t+1}^{v} = \alpha \tilde{W}^{T} f_{t+1}^{c} + (1-\alpha) y^{v} \end{cases},$$

y^{v} represents the initial scores of video node. For each event, we initialized its positive video nodes as score 1 and its negative video nodes as score -1. y^{c} represents the initial scores of concepts nodes and we initialized all the 346 concept nodes as score 0. f_t^{v} and f_t^{c} are the updated scores for video and concept nodes. $\tilde{W} = D_r^{-1/2} W D_c^{-1/2}$ is the normalized weight matrix, and D_r and D_c are the diagonal matrices with the row and column

sums of W in the diagonal. The propagation weight α was set to 0.5.

The propagation is stable once f_t^c has converged. The score of each concept node then indicates its relevance to the event. The concept node with strong connections to positive training video nodes will get high scores and the concept node with strong connections to negative training video nodes will get low scores. Table 1 shows three event-specific visual concept signatures. The left column is event name. The right column lists the Top-8 concepts for this event. (we ranked the concepts according to its score in f_t^c).

We observe that by using only the top 65 of 346 visual concepts, we can still achieve 90% performance in the MED task, so we restrict ourselves to these concepts, prune less discriminative concepts for recounting, and call the resulting ranking the event "signature", because it shows which concept are important.

Table 1. Topic-specific visual concept signatures computed by bipartite graph propagation (ranked according to f_t^c).

Events	Top 8 Concepts in signature
Flash mob gathering	Crowd, People_Marching, 3_Or_More_People, Demonstration_Or_Protest, Meeting, Cheering, Urban_Scenes, Walking
Parkour	Urban_Scenes, Building, Windows, Outdoor, Streets, Road, Walking_Running, Cityscape
Getting a vehicle unstuck	Car, Snow, Motorcycle, Outdoor, Landscape, Vehicle, Boat_Ship, Ground_Vehicles

3.2.2 Re-ranking using Ground Truth

In addition to the extracted SIN feature, we also know the event that the video belongs to. We can take this a-priori knowledge into account and refine the ranked list of visual concepts, so that concepts that humans think are important are preferred over other concepts. In our case, the ground truth for the "Parkour" event is {People_Marching, Demonstration_Or_Protest , 3_Or_More_People, Crowd, Adult, Cheering, Dancing, Walking, Joy, US_Flags, Urban_Scenes, Outdoor, Daytime_Outdoor, City, Streets, Vehicle, Road, Traffic, Meeting, Building, Politicians, Cityscape, Urban_Park, Trees, Road_Block}. This list was manually derived from a textual description of the "Parkour" event, however it could also be extracted automatically in the future, by using techniques similar to those discussed in the following section. Next, we re-rank the visual concepts with respect to the determined score as

$$score(i) = 1 / (r(i)/65 + mr(i))$$

where $r(i)$ is the rank of the remaining visual concepts filtered through the previous step and $mr(i)$ is the manually determined rank of the ith concept. For visualization, we can still determine a representative keyframe for each visual concept.

3.3 ASR Transcript Feature

We extract the words spoken in a video using ASR, as describd in [1]. We aim to identify the most relevant and informative words in the transcript with respect to the detected event. Conventionally, words with higher TFIDF score are considered more important. However, as we are observing around 60% word error rate, some words, which occurred only once and have relatively low TFIDF scores, can be highly related to the event and quit useful for TOMS. In addition, due to the presence of the ambient noises in the videos, many ASR transcripts include frequent words, which are incorrectly recognized and consequently they are not related to the event while they have relatively high TFIDF scores. To tackle this problem, we put more weights on words which are semantically related to the description of the detected event. We utilize integration of WordNet [5] and Wikipedia-based [8] similarities to measure the relatedness of each word to the event kit description of interest. Moreover, we determine unique words for each event (i.e. words occurred more frequently in a particular event) based on the given positive samples in the development data. Using the list of unique words for the event of interest, we assign higher weights on these unique words if they appear in the ASR transcript.

In conclusion, we can determine the score for each word in the ASR transcript as shown below and ranked them, accordingly.

$$score(i) = tf(i) * (1/\max_{j=1,...,n}(tf(j)) + WN\_Sim(i) + UNQ(i))$$

where i is the ith word and j is the jth word. The $tf(i)$ and $tf(j)$ are the term frequency of i and j. $WN\_Sim(i)$ is the maximum of the semantic similarities between the ith word in the ASR transcript and all words in the description of detected event. Note that we remove stop words and use the stemmer to convert every term to its original form before semantic similarity calculation. $UNQ(i)$ is 1 if the ith word in the ASR transcript is included in the unique word list of detected event. Otherwise it equals zero.

4. NATURAL LANGUAGE GENERATOR

We have implemented a Template-Based Natural Language Generation (NLG) System to generate written text about a given video. Figure 2 shows the general system architecture.

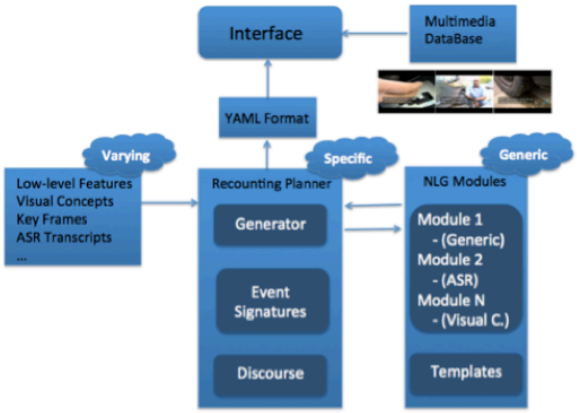

Figure 2. TOMS System Architecture: while the features vary for each video, the planner contains code that is specific for each topic, and the NLG modules/templates are completely generic.

The Recounting Planner's "Generator" receives the features extracted from the video (i.e. visual concepts with probabilities, ASR transcripts, etc.), and triggers several NLG modules to generate text using pre-defined, static templates. Currently, we have NLG modules that can deal with ASR output and visual concepts, generating one or more sentences each time they are called. A "generic" module generates text that does not directly refer to specific evidence. The planner calls other modules, such as the "activity" module and the "constrain" module, in order to generate high-level observations, which are technically generic,

but are often called for one or very few topics only.We are using the YAML markup language to abstract the recounting module from the user interface, and also for the communication between the individual modules. The user interface currently creates web pages, which can be shown in any browser, and also includes references to the original videos and keyframes, although we have so far allowed internal access only.In the following subsections we explains how each module works and what kind of result it will return.

4.1 Generic Module

The first module is the general module. It generates a general sentence talking about which topic this given video belongs to. Currently there are three natural language templates in the generic module:

This is a <Topic_Name> event.

The video shows the event of < Topic _Name>.

This video is about < Topic _Name>.

Using the label given by the detection part, we fill in the blank with the name of the event. For example, if the label is "Birthday_Party", we just fit this event name in one of the three templates (randomly picked) and compose a sentence like: "This is a Birthday_Party event." The use of several templates reduces the monotonicity of the recounting, while preserving accuracy.

4.2 Visual Concept Module

The visual concept module generates several sentences talking about the objects and scenes that are observed in the video. The input feature is a ranked list of the visual concepts, together with their confidence scores. The visual concept module executes the algorithm described in Section 3.2, to determine which features to mention in the recounting for this specific video, and this event.

After re-ranking the visual concepts, we pick the top 5 percent concepts as the video's visual concepts and use them to generate recounting sentences. These top 5 percent visual concepts are then compared with the topic signatures (see Table 1 for examples) signatures and divided into two subsets: the positive subset and negative subset. If a concept in this video can be found in the event's "most relevant" signatures, which are the top visual concepts in the event's signature, then this concept is assigned to "positive" subset; if a concept we detected from the video exists in the event's "least relevant" signature list (the last 50 visual concepts in the event's signature list), we regard it as a "negative" visual concept. We use the "positive" subset of visual concepts to generate one to three recounting sentences, and use the "negative" visual concepts to generate one to two recounting sentences.

In the three sentences that address the positive visual concepts, we set two thresholds to separate the "most relevant" visual concepts according to different confidence values. If the confidence value is larger than 0.6, we use the following template to generate a sentence:

We saw <List_of_Visual_Concepts> in the video.

If a visual concept's confidence value is less that 0.6, we employ the template:

We <adv> saw <List_of_Visual_Concepts> in the video.

The adverb here has two different values: probable and possible. If the confidence value of a visual concept is less than 0.6 but higher than 0.3, we choose the preposition "probable". If the confidence value is less than 0.3, we just use the preposition "possible" because our system is not very sure about this visual concept. We introduced this distinction in response to initial user tests, as described in the next section of this paper.

An example recounting text generated from the Visual Concept Module could be like:

We saw Body_Parts in the video. We probably saw Indoor and Room in the video. We possibly saw 3_Or_More_People, Food and Joy in the video.

While it is clear that the quality of the text can be improved (we could for example map "Body_Parts" to "body parts" on the screen), we retain this format for debugging purposes for now.

4.3 Module for Text Concepts

The format of the ASR Transcription features entering the TOMS system is a list of high-level semantic words (like car, open, tool, etc.). The scoring and ranking method for these features have been described in Section 3.3.

With the ranked ASR Transcription list, some templates are generated to express these transcriptions in natural language. The template in this module is similar to the visual concept module:

We <adv> heard the words <List_of_ASR Transcriptions> from the video.

If one word has very high confidence and is very related to the event, we just omit the adverb. If the system is not that sure about whether it heard the word in the video, we put "*probably*" as the adverb here to generate a sentence like:

We probably heard the words glass, clean and hand from the video.

Again, both types of sentences can be produced, if required, and the confidence values have been set empirically for now.

4.4 Activity Module

The "activity" module implements a grammar-based algorithm, which attempts to generate more relevant and complex sentences from certain, frequently observed combinations of visual concepts, than the baseline visual concept module.

In order to address "activities" in the video, we manually labeled all 364 visual concepts with a tag, defining the category of this concept. Currently, we are using 4 kinds of tags: Subject, Activity, Object, and Location. "Subject" refers to the concepts that can be subjects in the sentence, like "Adult", "3_Or_More_People" and "Driver". "Activity" contains the visual concepts that explicitly show an activity: "Bicycling", "Car_Racing" and "Dancing". "Object" means concepts that are typically referred to as an object, such as "Cell_Phones", "Chair", "Factory". The "Location" tag is given to the visual concepts that are locations or scenes: "Doorway", "Fields", "Forests". Again, we implemented several templates that can be used with concepts that are labeled with these tags, for example:

We detect <Object> <Activity> (<Object>) in <Location>

to generate sentences about concurrent activities that are happening in the video. One example result given by the Activity Module could be:

In this video we detected Adult Talking in Kitchen.

In the future, we plan to employ statistical language models and parsers to improve the fluency of the output. At present, we do not require that these concepts be detected at the same time in the

video, as we have not found examples in our database violating this condition.

4.5 Concept-Constrain Module

This is an additional module that we only use in some videos and topics. The activity during events such as "Parade" and "Flash_Mob_Gathering" is supposed to happen outdoors rather than indoors. This could be regarded as a constraint for this event. Cases where videos that were labeled as "Parade" or "Flash_Mob_Gathering" events show high confidence measures for "indoor" are addressed by generating a sentence such as::

The Parade event is more likely to be an outdoor event, but we believe this video is indoors.

Similarly to the presentation of unexpected visual concepts in the initial part of the recounting, this module generates specialized sentences for unexpected combinations of features, which can be an important detail for the understanding of a video.

Finally, the interface of the current Topic-Oriented Multimedia Summarization System is shown in Figure 3.

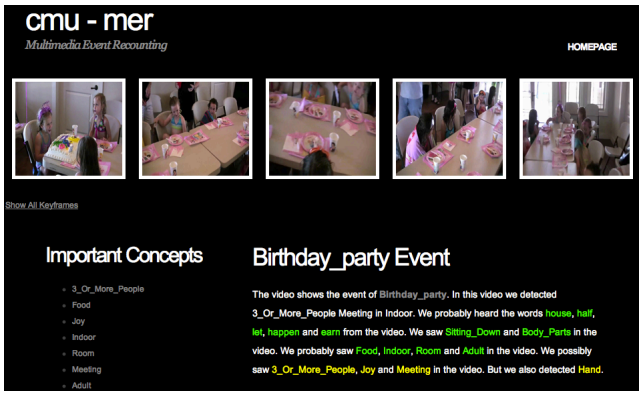

Figure 3. The User Interface of our current TOMS System. For diagnostic purposes, visual and audio concepts are color coded, and the keyframes related to important concepts can be selected directly. The interface is realized as a web page.

5. EXPERIMENTS

Using the output of our current TOMS system and demonstrator, we conducted two pilot studies to investigate the following question:

To what extent can machine-generated recounting summaries, compared with human-generated ones, help people recover information from a multimedia material?

Specifically, we ran a study looking at the indicative effectiveness of a textual recounting: how well can users identify which event is indicated in the recounting. We also ran a study looking at informational effectiveness, i.e. can a user identify which video in the same event class matches the given text recounting.

5.1 Experimental Paradigm

We compare the recounting passages generated by our TOMS System with human created recounting summaries in information-recovery tasks to show how effective the system can be in accomplishing the recounting goals of indicativeness ("is this video an example of an event?") and informativeness ("what is in this specific video?"). The information-recovery tasks include Event Selection and Video Selection tasks, which allow us to measure summarization quality progress and optimize the system.

5.2 Dataset Preparation

We first collect a set of 20 recounting text passages for 20 different videos in the dataset. Among these 20 recounting passages, 10 passages were automatically generated from the TOMS system and the other 10 passages were written by a person. These 20 passages are divided into four groups: Group 1 and Group 2 were designed for the event selection task, and Group 3 and Group 4 were defined for the video selection task. The composition of the 4 groups is as the following:

- Group 1 contains five recounting passages, which are generated by TOMS system. For each of the 5 recountings exist *one label that describes the event correctly, and two confusing labels* that are associated with the recounting;
- Group 2 contains 5 recounting passages, which are generated by human editors. For each of the five recountings exist *one label that describes the event correctly, and two confusing labels* that are associated with the recounting;
- Group 3 also contains five recountings, which are generated by TOMS system. For each recounting, we show *one video that is the correct fit, and two confusing videos* associated with the recounting;
- Group 4 still contains 5 recounting passages, which are generated by human editors. For each recounting, we show *one video that is the correct fit, and two confusing videos* associated with the recounting;

The preparation of the dataset is shown in Table 2.

Table 2. The composition of the experimental dataset

Generated by	Event Selection Task	Video Selection Task
TOMS	5 samples	5 samples
Human	5 samples	5 samples

5.3 Pilot Study 1: Event Selection

Test subjects see ten recounting passages, five of which are generated by TOMS and five by human editors. The human editor generated the text with no knowledge of the events, i.e., it was an informational summary only. The TOMS method generated evidence in favor of the video represented by the recounting belonging to one of the ten TRECVID 2011 MED events. Three labels were displayed to the subject with each recounting, one correct answer event label and two confusing event labels, as shown in Figure 4.

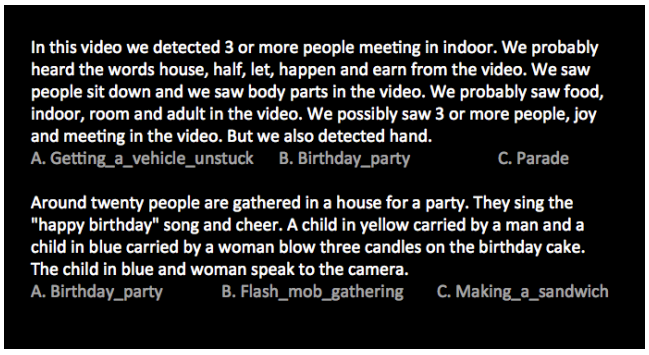

Figure 4. TOMS Event Selection Task Interface.

5.4 Pilot Study 2: Video Selection

Test subjects see ten recounting passages, five of which are generated by TOMS and five by human editors. The human editor generated the text with no knowledge of the events, i.e. it was an informational summary only. The TOMS method generated evidence in favor of the video represented by the recounting belonging to one of the ten TRECVID 2011 MED events. In this study, event labels are not used. Rather, the subject is offered three videos, and asked which of the videos the recounting represents. The task is made more difficult for an indicative summary in that all three videos show the same event, e.g., they all show "birthday party" for one recounting, and all show "parkour" for another. This pilot test stresses the capability of an indicative summary like TOMS being able to also act as an informational summary representing a specific video.

5.5 Experimental Procedure and Results

We invited ten people to participate in our pilot studies, based on easy access to our labs (hence the "pilot" studies rather than more formal follow-up work). Nine were male, all were experienced computer users who watch computer-delivered video and read English text well. Each subject was introduced to the recounting idea (represent a video with text), and the two pilot studies. Each subject completed all twenty judgments (ten each for the two pilot studies).

The overall result is shown in the Figure 5 for the ten participants (P1-P10, abscissa), with the maximum number of correct answers per study being ten. The event selection tasks, i.e., the indicative aspect of recounting, are much simpler in that the events are very different from each other, everything is represented as text (Figure 4), and TOMS is geared toward providing evidence in support of membership for one of the listed event classes. For the video selection problems, the answer video and other two videos are always from the same event class, making it harder to choose the right one.

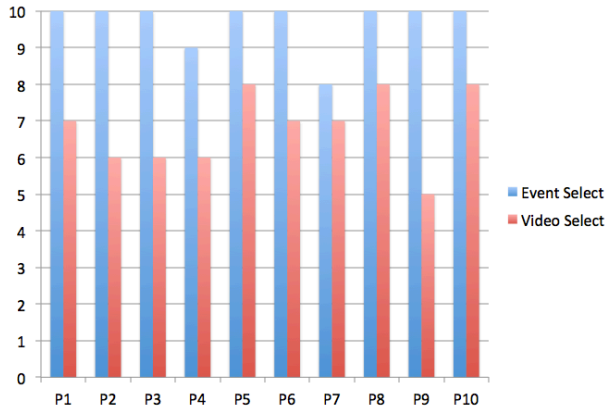

Figure 5. Overall Result of Pilot User Study: it is harder to determine the actual video belonging to a recounting, rather than just the topic of the recounting.

For event selection, the performance is shown in Figure 6.

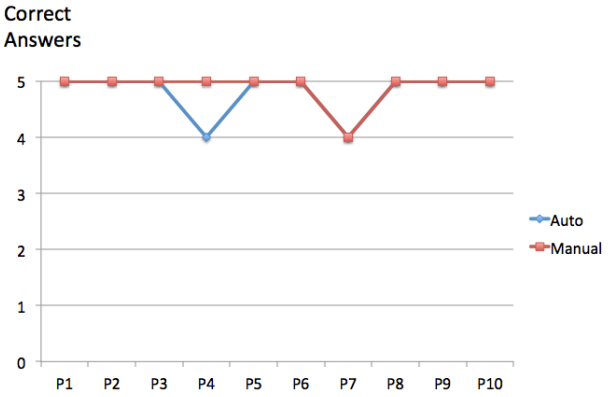

Figure 6. Comparison of performance of automatically-generated and manually-generated summaries in Event Selection Tasks.

Both auto-generated and manually-generated recounting summaries have excellent performance in the event selection tasks. Only one question is incorrectly answered in the manually-generated problems, with two questions having wrong answers in the auto-generated problems.

The results for the video selection tasks are shown in Figure 7.

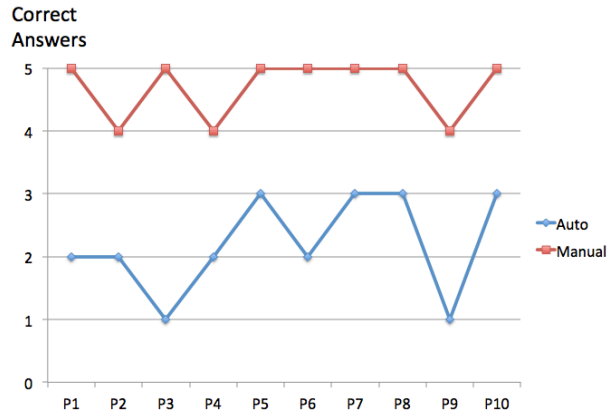

Figure 7. Comparison of performance of automatically-generated and manually-generated summaries in Video Selection Tasks.

The manually-generated summaries outperform the auto-generated summaries in the video selection tasks. The average score of the manually-generated problems is 4.7, while the average score of the auto-generated problems is 2.1. This is partly because the current TOMS system can only generated relatively more general features and concepts of the video, while the human editor can grasp more detailed and specific features from the video. The TOMS system was architected to provide evidence indicating membership in one of the MED event classes. The human summary was completely informational, generated without regard to any event classes and hence often included unusual features such as the display of foreign text overlays or particular noises. These detailed features are often very helpful in directing users to choose the right answer.

From the user study we also found that temporal construct information is very helpful for participants in quickly making the

right choice. For example, in the video selection task, one of the human generated recounting passages starts with "A movie title with words in red…" The user can immediately distinguish the answer video from the other two videos based on this information. Moreover, from the human generated summaries we also found that, specific details of the video can help a lot in making the choice. (e.g. "red hat", "birthday cake", "a tall policeman" etc.) These obained experience will help us to make improvements to the TOMS system in next steps.

The aim of these pilot user studies is to propose a method that can be used for evaluating the performance of multimedia summarization systems directed toward indicative and informative purposes. It also provides some preliminary results when focused only on text representation. This methodology is derived from the principles of user-centric design, and can be applied to many Topic-Oriented Multimedia Summarization systems. The current demonstrator platform can be made available for crowd-sourcing experiments and demonstration purposes on the web, so our initial experimental results can be used as benchmarks in future work.

6. CONCLUSION

In this paper we first reviewed recent and ongoing research in the area of multimedia summarization. We motivated the need to go beyond browsing based retrieval paradigms, and defined the task of Topic-Oriented Multimedia Summarization (TOMS). We differentiated our work from prior systems in that we are investigating static summaries with a text component, i.e., ones that can be viewed all at once rather than playable video gists or skims that have a temporal element.Our TOMS system is presented, which is capable of generating text-based recountings for videos belonging to one of ten MED events in a database of (currently) 1200 hours. Our automatic system includes and fuses state-of-the-art audio and video features, and attempts to explain why a certain video is assigned to a certain event, or how videos belonging to the same event differ. One method for evaluating multimedia summarization systems is proposed, followed by pilot user studies and preliminary results.

In future work, we will incorporate more features to guide the recounting, further improve our system using a user centric design process, and try to scale and automate the evaluation process, addressing both indicative and informative aspects of recounting.

7. ACKNOWLEDGMENTS

Removed for anonymous review.

8. REFERENCES

[1] [Removed for anonymous review] *In Proc. TRECVID2011, NIST.*

[2] [Removed for anonymous review] *In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA.*

[3] Michael G. Christel. Evaluation and User Studies with Respect to Video Summarization and Browsing. *In Proceeding of "Multimedia Content Analysis, Management, and Retrieval", part of the IS&T/SPIE Symposium on Electronic Imaging, San Jose, CA, January 17-19, 2006.*

[4] Michael G. Christel. Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation. *San Rafael, CA: Morgan and Claypool Publishers, 2009.*

[5] Quang Do, Dan Roth, Mark Sammons, Yuancheng Tu and V.G. Vinod Vydiswaran. Robust, Light-weight Approaches to compute Lexical Similarity. *Computer Science Research and Technical Reports, University of Illinois-2009.*

[6] Alexander G. Hauptmann, Michael G. Christel, Wei-Hao Lin, Bryan Maher, Jun Yang, Robert V. Baron, Guang Xiang. Clever clustering vs. simple speed-up for summarizing rushes. *TVS '07 Proc TRECVID Video Summarization Workshop, 2007*.

[7] Chiori Hori and Sadaoki Furui. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems E87-D(1):15–25. 2004*.

[8] Peter Kolb. Experiments on the difference between semantic similarity and relatedness. *In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09, Odense, Denmark, May 2009.*

[9] Brian Langner and Alan Black (2009) MOUNTAIN: A Translation-Based Approach to Natural Language Generation for Dialog Systems, *In Proceeding of IWSDS 2009, Irsee, Germany.*

[10] Yingbo Li, Bernardo Merialdo. Multi-video Summarization Based on AV-MMR. In *Proc. 2010 Int'l Workshop on Content-Based Multimedia Indexing*, 1-6.

[11] Gary Marchionini, Yaxiao Song, and Robert Ferrell. Multimedia Surrogates for Video Gisting: Toward Combining Spoken Words and Imagery. *Information Processing & Management* 45(6), 2009, 615-630.

[12] NIST, Information Technology Laboratory. "2011 TRECVID Multimedia Event Detection Track," http://www.nist.gov/itl/iad/mig/med11.cfm. Last accessed January 28, 2012

[13] Ani Nenkova. Summarization evaluation for text and speech: issues and approaches. *In Proc. INTERSPEECH, Pittsburgh, PA; USA. ISCA, 2006.*

[14] Ba Tu Truong and Svetha Venkatesh. Video Abstraction: A Systematic Review and Classification. *ACM Trans. Multimedia Computing, Communications, and Applications (TOMCCAP) 3(1), 2007, 1-37*.

[15] Chun Chet Tan, Yu-Gang Jiang, Chong-Wah Ngo. Towards Textually Describing Complex Video Contents with Audio-Visual Concept Classifiers. *In Proceeding of ACM Multimedia 2011, Scottsdale, Arizona, USA.*

[16] Yoshitaka Ushiku, Tatsuya Harada, Yasuo Kuniyashi. Understanding Images with Natural Sentences. *In Proceeding of ACM Multimedia 2011, Scottsdale, Arizona, USA.*

[17] Robin Valenza, Tony Robinson, Marianne Hickey, and Roger Tucker, "Summarization of spoken audio through information extraction," *In Proceeding of ESCA Workshop on Accessing Information in Spoken Audio, 1999, pp.111–116.*