Int J Multimed Info Retr (2013) 2:131–144
DOI 10.1007/s13735-012-0028-y

REGULAR PAPER

Beyond audio and video retrieval: topic-oriented multimedia summarization

Florian Metze · Duo Ding · Ehsan Younessian · Alexander Hauptmann

Received: 15 October 2012 / Accepted: 13 November 2012 / Published online: 4 January 2013

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 (NLG): given a set of automatically extracted features from a video, a TOMS system will automatically generate a paragraph of natural language, which summarizes the important information in a video belonging to a certain topic, and for example provides explanations for why a video was matched and retrieved. Possible features include visual semantic concepts, objects, and actions, environmental sounds, and transcripts from automatic speech recognition (ASR). We see this as a first step towards systems that will be able to discriminate 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 various visual concept features, environmental sounds and ASR transcription features from a given video, and develop a template-based 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.

Keywords Multimedia summarization · Event detection and recounting · Natural language generation

F. Metze (✉) · D. Ding · E. Younessian · A. Hauptmann
School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
e-mail: fmetze@cs.cmu.edu

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 in this paper we will argue that natural language can play a big role in organizing this information.

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 temporal aspects will be taken into account, which can give users a unified, coherent explanation of what is happening in a video. Because it not just returns a list of audio/video concepts 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, the "features", that can be observed in a video. We do this on a dataset of videos with topic labels [26] (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 observations that define the topic. These observations can eventually also be defined as plain text, by giving a textual description of important objects and actions that make up each topic. An earlier version of this system was described in [8].

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. In the beginning, we heard wedding music, and later people talking about the beautiful bride.

Our TOMS system uses the same features we also use for automatically classifying videos into these topics, or events [3,17], and was first demonstrated in [7]. In this paper, we first present a set of exploratory experiments to generate text from detected features, and to select salient features and feature groups. We are currently focusing on text generation for the two most salient semantic types of features, out of four, which our group currently uses for video classification:

- 346 "SIN" [28] video semantic concepts ("vehicle", "animal", "body parts", etc.) and ObjectBank [27] features describe visual information, while
- English ASR output [14] captures the information in the audio signal.
- Audio semantic concepts ("noisemes" [14], such as "engine", "noise", or "music") describe more general audio scenes, and were added to analyze the saliency of different feature types. In future work, we plan to also integrate
- OCR features into the summarization process.

Eventually "actions" and other complex semantic temporal and spatial 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 at this point.

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), as used in the MOUNTAIN system [16].

The paper is organized as follows. Section 2 reviews related work in the area of audio and video summarization. In Sect. 3, we briefly present the multimedia features that we extract, and use for classification. Section 4 describes the NLG system. Our exploratory user study on text generation is described in Sect. 5, while Sect. 6 describes our initial experiments on the salience of various feature types, and ways to select the information that needs to be presented. Finally, Sect. 7 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. [35] 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 [13] or sentences from the original data, or by synthesizing new representations from the original data [20]. The second case is of course more difficult and 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 [1], and how speech summarization is possible based on non-textual features alone [4,11]. 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 [18,24].

Of course all audio summarization work is tied closely to progress in our understanding of basic speech-to-text and speaker diarization algorithms, as well as audio event recognition or "acoustic scene analysis" [21].

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 [33] 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. [5,6,12] 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 describe the merits of discount usability techniques for iterative improvement and evaluation. Christel et al. 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. Interestingly, most work finds no correlation between performance and satisfaction measures [37]. 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, which were being used to help the browsing process, for example in the Open Video Archive [22]. Video summaries had a temporal aspect in terms of playable audio-visual material [12,22], or were even seen as new "narratives" [38], a form of storytelling, without conversions of modality. The informative summary for a video exploiting both audio and video information was improved with a maximal marginal relevance algorithm working across video genres [18]. Work has been performed towards automating content-based evaluation of summaries, particularly of BBC rushes, in the context of TrecVID [10]. The similarity to work in text summarization is recognized by naming a proposed evaluation metric for video summaries VERT [19]. Work continues to understand and improve the user interface to video summarization, for example by providing an opportunity to create videos that match a story given in text [29], or by transforming 2D videos into a 3D cube, which can be navigated [25].

In this work, our focus is on static presentation, emphasizing textual summaries as done by Ushiku et al. [34]. 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. [32] have proposed utilizing audio-visual concept classifiers to generate textual descriptions of video content. 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. To evaluate, 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.

Again, work in basic video retrieval contributes to video summarization, see for example [30] or the NIST's TrecVID evaluation campaign [28] for discussion of the state-of-the-art. Given that information to be included in the summary can typically be drawn from the large pool of data generated during retrieval, feature selection strategies become a

pressing problem: what to include in a summary, and why? This questions has also been recognized by other researchers, for example [37], who investigated different ways to present visual information. Our work is different by using text as the only, "unified" presentation modality. A similar approach was taken by Song et al. [31], who investigated "eye-catching and ear-catching" content of instructional videos, but had humans create summaries, while our process is fully automatic. Their evaluators judged for example that a segment in which a single person speaks is a good candidate to be included in a summary, which seems reasonable for a tutorial video.

Another avenue that we believe text-based video summarization lends itself well to, even though we do not explicit this aspect in this paper, is multi-document summarization [36].

3 Multi-media feature definition and extraction

In this paper, each video is labeled as one of ten events (topics) from the TRECVID 2011 Multimedia Event Detection (MED) task and database [26]:

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

In more recent work, we have extended our system to cover five more events from the 2012 task [23]. Currently, about 1,000 videos have manual (reference) topic labels, while about 150,000 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 15 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 not only for reference topic labels but also for automatically generated topic labels, which might well be wrong.

3.1 Video-level SIN feature

We employ the visual concept detector to index all extracted keyframes from a given video. For each keyframe, we calculate scores for each of the 346 visual concepts. These visual concept detectors are SVM classifiers trained over the SIN task in TRECVID 2011 using MOSIFT and CSIFT features to describe keyframes [3]. To determine the video-level semantic indexing, we simply take the average of the keyframe-level SIN for all keyframes within a video. We evaluated different ways to determine the video-level SIN representation such as taking the max, median and mean of the keyframe-level SIN. Our experimental results show the superiority of taking the average to merge the keyframe-level SIN to generate video-level representations.

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" and "indoor" 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 (often around 0.17), many of them are essentially unreasonable, and should not be used for sorting directly.

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.

Let $G = (V, C, E, W)$ be a 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_j. The propagation process in the graph can be written as:

$$f^c_{t+1} = \alpha \tilde{W}^T f^v_t + (1-\alpha) y^c$$
$$f^v_{t+1} = \alpha \tilde{W}^T f^c_{t+1} + (1-\alpha) y^v$$

y^v represents the initial scores of the video nodes. For each event, we initialized its positive video nodes with $w^v = 1$, and its negative video nodes with $w^v = -1$. y^c represents the initial scores of concepts nodes and we initialized all the 346 concept nodes with $y^c = 0$. f^v_t and f^c_t are the updated scores for video and concept nodes. $\tilde{W} = D_r^{-0.5} W D_c^{-0.5}$ 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^c_t 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 topic-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^c_t).

Taking the minimum of determined ranks in different events for a particular visual concept, we determine its global rank. For instance, "car" is ranked as the Top-1 visual concept since it is Top-1 for the second and the fourth event. 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 ranking result the event "signature", because it shows which concepts are important.

Table 1 Topic-specific visual concept signatures computed by bipartite graph propagation (ranked according to f^c_t)

Event	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 human-generated list of relevant concepts

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 relevant are preferred over other concepts. In our case, the list of relevant concepts for the "Parade" 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 "Parade" event, provided together with the videos. It could also be extracted automatically in the future, by using techniques similar to those discussed in the previous section. Visual concepts are re-ranked with respect to

$$\text{Score}_V(c) = \frac{1}{R(c)/65 + R_M(c)}$$

where c and $R(c)$ refer to the remaining visual concepts in the signature, determined in the previous section, and their rank, respectively. $R_M(c)$ is the rank of c in the human-generated list of relevant concepts. After the re-ranking of visual concepts, for visualization, we determine a representative keyframe for each visual concept to be the one that has the maximum score for the corresponding visual concept.

3.3 ASR transcript feature

We extract the words spoken in a video using ASR, as described in [3]. 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 in a video and have relatively low TFIDF scores, can be highly related to the event and quite 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 the integration of WordNet [9] and Wikipedia-based [15] 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

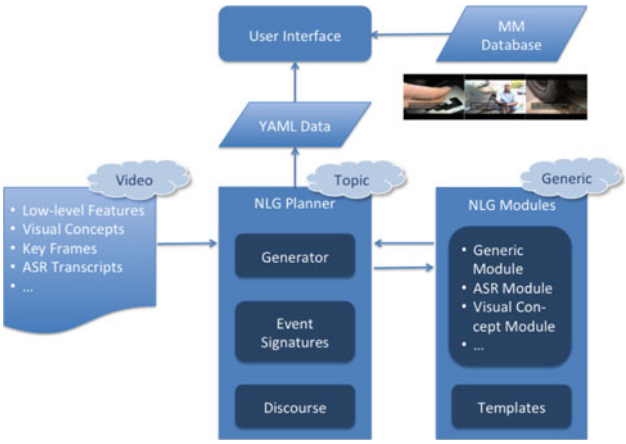

Fig. 1 TOMS system architecture: while the features vary for each video, the planner contains code that is specific for each topic

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 using:

$$\text{Score}_T(t_i) = \text{TF}(t_i) * \left(\frac{1}{\max_{j=1,\ldots,n} \text{TF}(t_j)} + \text{WNSIM}(t_i) + \text{UNQ}(t_i) \right)$$

where t_i and t_j are the ith and the jth terms in the ASR transcript of a video, respectively. $\text{TF}(t_i)$ is the term frequency of t_i. $\text{WNSIM}(t_i)$ is the maximum of the semantic similarities between t_i and all words in the description of detected event. n is the number of terms in the ASR transcript of the video. Note that we remove stop words and use the stemmer to convert every term to its original form before semantic similarity calculation. $\text{UNQ}(t_i)$ is 1 if t_i is included in the unique word list of detected event. Otherwise it equals zero.

4 Natural language generator

We have implemented a template-based NLG system to generate written text about a given video. Figure 1 shows the general system architecture.

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 explain 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 Sect. 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) and divided into two subsets: the positive subset and the 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 the "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 than 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 Sect. 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> in 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 in 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 four 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. Similar 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 TOMS System is shown in Fig. 2.

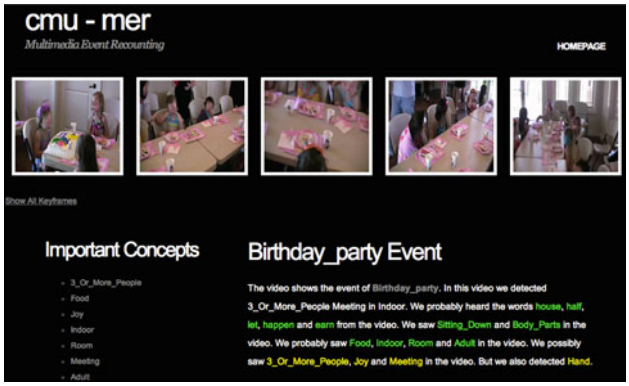

Fig. 2 The user interface of our current TOMS system. For diagnostic purposes, visual and audio concepts are color coded, and the keyframes related to important concepts can be selected directly. The interface is realized as a web page

5 Text generation pilot 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:

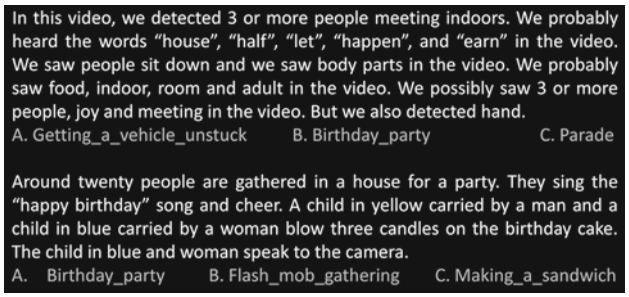

Fig. 3 TOMS event selection task interface

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 four groups is as follows:

- Group 1 contains five recounting passages, which are generated by TOMS system. 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 2 contains 5 recounting passages, which are generated by human editors. For each of the five recountings exists 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 five 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.

For both tasks, we had five samples generated by humans, and five samples generated by our TOMS system.

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 Fig. 3.

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 humans

were told to generate a summary of a video informing a reader about what might be in that video, but without any further limiting or guiding context. The human was not told to write up why a given video belongs to the birthday party event (i.e., the human was not told to generate an indicative summary). The TOMS method generated evidence in favor of the video represented by the recounting belonging to one of the ten TRECVID 2011 MED events. So it is set up to do an indicative summary as to why the video belongs to the class of videos for a given event kit (i.e., indicate why a video belongs to a certain class). 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 subjects to participate in our pilot studies, mostly students at Carnegie Mellon University, to which we had easy access. 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 Fig. 4 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 (Fig. 3), 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 the other two videos are always from the same event class, making it harder to choose the right one.

The performance on the event selection task is shown in Fig. 5. Both auto-generated and manually generated recounting summaries have excellent performance in the event selection task. Only one question is answered incorrectly for the manually generated recounting, while two questions trigger wrong answers for the auto-generated recountings.

The results for the video selection tasks are shown in Fig. 6. The manually generated summaries outperform the auto-generated summaries in the video selection tasks. The average number of correct answers for the manually generated recounting per test subject is 4.7, while the number for auto-generated recountings is 2.1. This is partly because the current TOMS system can only render quite general features

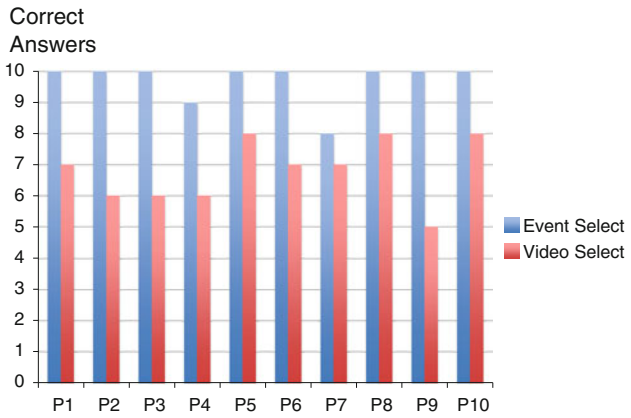

Fig. 4 Overall result of pilot user study: it is harder to determine the actual video belonging to a recounting, rather than just the topic of the recounting

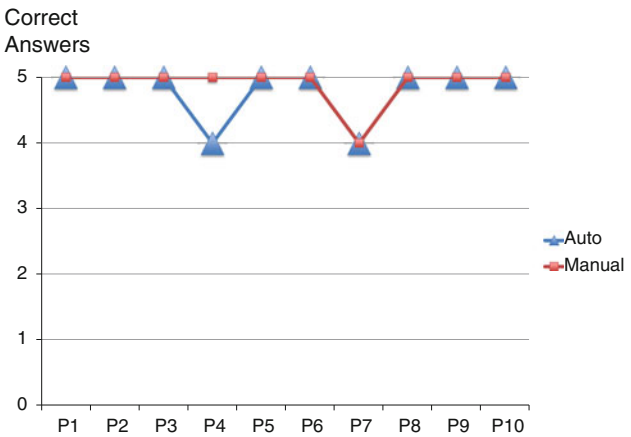

Fig. 5 Comparison of performance of automatically and manually generated summaries in event selection tasks

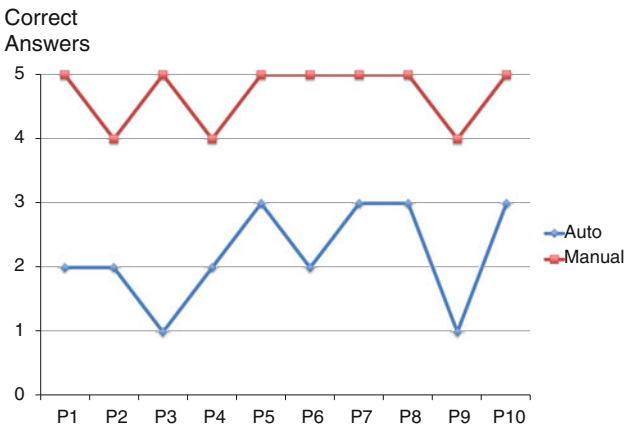

Fig. 6 Comparison of performance of automatically and manually generated summaries in video selection tasks

and concepts of the video, while a human author can describe more detailed and specific characteristics of 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 characteristics, e.g. the display of foreign text overlays or particular noises. These details are often very helpful for users to choose the right answer.

From the user study we also found that temporal information is very helpful for participants. 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 "correct" video from the other two videos based on this information, because it is the only one that starts with a title screen in red. Moreover, from the human-generated summaries we also found that specific details of the video can help a lot in making a choice (e.g. "red hat", "birthday cake", "a tall policeman", etc.).

These observations will help us make improvements to the TOMS system in the future. In the next section, we present a pilot study to investigate how information presentation can be evaluated and improved.

6 Salient feature selection pilot experiments

When designing text-based video summarization, multiple kinds of features can be taken into account, and need to be presented suitably in the common text modality. But how important is each feature for recognizing and describing its content? Will the users be confused by some of the features? How many top results (values) for each kind of feature should be presented? In what order?

Using the features of our current TOMS system and demonstrator, we conducted a third pilot study to investigate the following question:

> Given a set of different types of features, which of these will a user pay more attention to, when included in a text-based summarization of a video?

6.1 Experimental paradigm

The underlying idea is to treat the user's attention as a limited resource, and force the participants in the study to take decisions, in which they prefer one set of values of a certain feature over another value, or show that a particular feature goes unnoticed, no matter the feature values. Of course it would be possible, and ultimately desirable, to evaluate features directly on their influence on the performance in a task at hand, but such evaluations are necessarily very narrow in their focus, and expensive. The proposed approach on the other hand is very general, and allows us to extract a maximum of diagnostic information about the quality of the extracted features, and the way the features are presented in text form using a simple unbiased user interface.

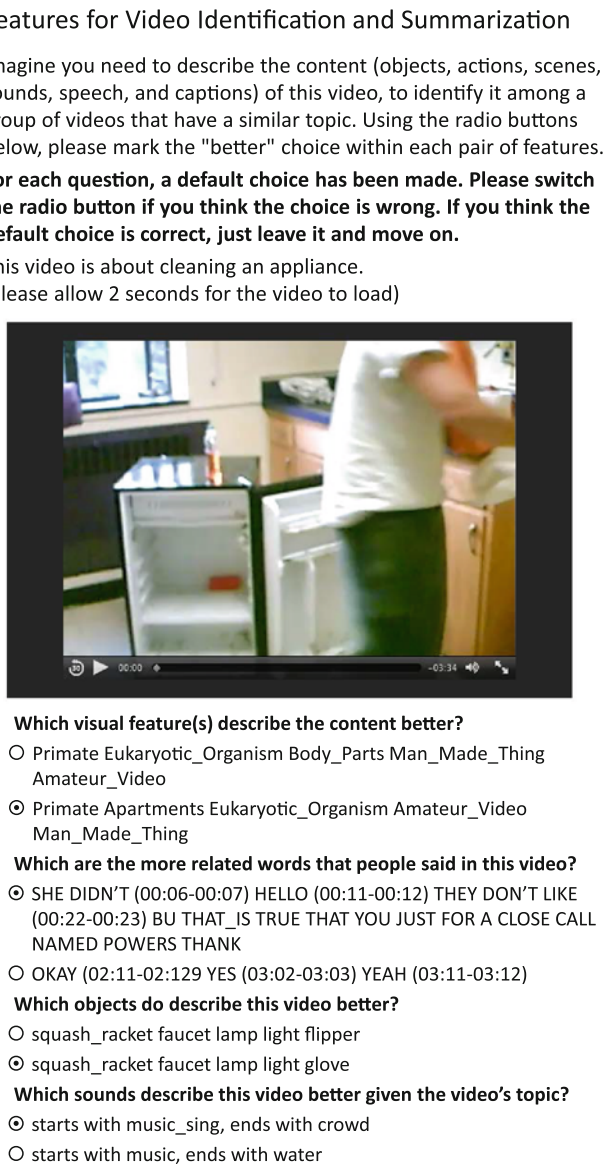

Fig. 7 TOMS feature selection task interface

We implemented a web page-based user interface for this experiment, shown in Fig. 7. Because the focus is on identifying useful features to include in the summarization, rather than judging the quality of text generation, we did not use the NLG components described earlier, but present "raw" feature values instead.

The participants are asked to first watch a video, and then answer several questions by clicking on radio buttons. The user's "mental model" is a need to identify the "best" answer to each question, in a video selection setting. Each question represents a certain feature type, i.e. the first question covers SIN visual concepts, the second question presents different ASR transcript features, etc. One of the two possible answers to each question was pre-selected for each participant. Participants were "paired" so that one participant had those items

pre-selected, that another participant did not see pre-selected. These assignments were done randomly. In this setup, we expect to be able to categorize the features according to the following observations:

- "Good" or "important" features: if the participant thinks the pre-selection is correct, he or she will just leave the pre-selection untouched. In this case, the "paired" user will switch the answer, and the "accuracy" of the final selection will be well above the 50 % chance level.
- "Irrelevant" or "uninformative" features: users will not switch answers, irrespective of the pre-selection, i.e. both the "correct" and the "incorrect" answers will tend to remain selected, when pre-selected, and result in about 50 % of the answers being "correct".
- "Confusing" features: users tend to switch the "correct" as well as the "wrong" answer to each question, or select the "wrong" answer with more than 50 % probability.

While the users do not actively perform a task in this setup, and we did not enforce any time or other constraints, users will tend to focus their attention on the feature type they find most useful, and ignore the others. The experiment should therefore tell us, which features users care about, which ones do not interest them, and which ones the current interface presents well, given the automatic extraction and presentation of features.

6.2 Dataset preparation

The dataset in the user study includes six videos from MED12's "Cleaning_an_appliance" event. For each video, we extract four different kinds of features: SIN visual concepts, ASR transcripts, ObjectBank results, and "Noiseme" semantic audio concepts. Two of them are from the "visual" modality, while the other two features come from the "acoustic" modality. SIN visual concepts represent people, actions, or scenes that are observed in the video, for example "outdoor", "road" or "3_or_more_people" (cf. Sect. 3.1). The ObjectBank detectors output labels such as "cow", "car", or "airplane" [27]. ASR output was segmented and filtered for noises before printing, and "Noisemes" represented 42 classes such as "music" and "engine noise".

Similar to the previous experiment, we recruited 16 internal participants who knew about the overall goal of the TOMS system, but were not told about the exact goal of this study before participation. In particular, they did not know what the "expected" values of the features were for the given event.

The textual values of the answers were automatically generated using our 2012 TRECVID MED and MER systems. The "correct" answer to each question was taken from the actual system output for the corresponding video, while the "competing" answer came from a randomly selected different video in the test set for this experiment. In this sense, the setup is similar to the "Video Selection" task presented earlier, and the outcome will reflect the suitability of a feature given the performance of the current feature extraction and presentation: if a feature cannot be extracted reliably, or is presented in an unsuitable format, we expect that the participants will not be able to identify the "correct" answer reliably.

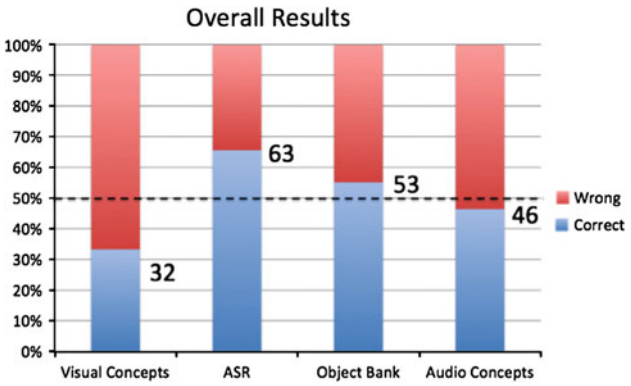

Fig. 8 Accuracy of participants' classification of feature values for the four feature types. Chance level is at 50 %. Participants can identify the "correct" ASR feature correctly in 63 % of cases (statistically significant at the 0.1 % level), while the SIN visual concepts feature is confusing (also at 0.1 % level)

6.3 Pilot study 3: feature selection experiment and results

From the log files collected during completion of the web forms, we extracted two types of measurements. The classification accuracy measures how many questions the participants answered correctly, i.e. how often the automatically extracted feature extracted from the reference video was preferred over the feature value extracted from a competing video. The second measurement is the "switch rate", which describes the percentage of answers which a participant switched from the default, pre-selected response.[1]

Given 16 participants who each labeled 6 videos, we have 96 data points in total, each describing 4 feature types. Figure 8 shows the accuracy of the participants' response, while Fig. 9 shows the amount and type of switching the participants exhibited.

The results show that, in this implementation and presentation, only the ASR features could probably provide useful information to a user. Given two competing ASR feature values, participants are able to identify the correct one, and will typically correct the wrongly pre-selected value, even when not required to do so. "Noiseme" and "ObjectBank" features

[1] Our current experimental setup does not allow us to measure "click rate", which could be used to see if participants switched an answer multiple times, switching back and forth the response, possibly indicating confusion.

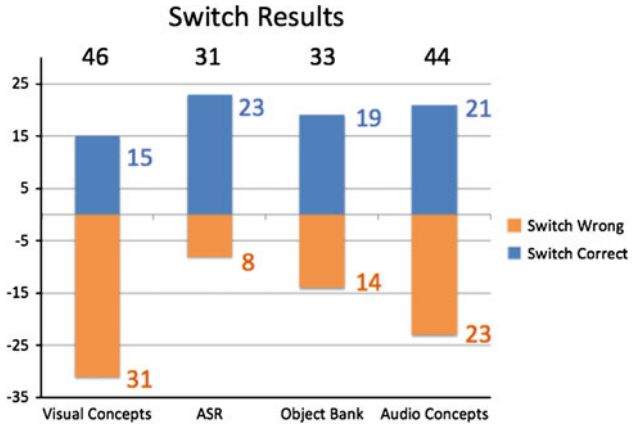

Fig. 9 Degree and direction of participants' "switching" of pre-selected answers. 31 visual concepts (of 48) were switched from the correctly pre-selected answer to a wrong answer, while only 15 were switched from the wrong answer to the correct one. The total amount of switching was also greatest for the SIN visual concepts (46 switches), while ASR was only switched in 31 cases (effectively correcting 15 answers)

are selected roughly at chance level, and while almost half the users switch the "ObjectBank" feature, and a third of the users switch the "noiseme" feature, the resulting accuracy is not better than chance at the 5 % level. It is interesting to note that the visual "ObjectBank" feature appears to be less prominent than the "noiseme" feature, which participants corrected more often (as often, in fact, as the ASR features). The "SIN" features, however, with the current confidence-based selection algorithm, are confusing to the user, because their accuracy is well below chance. Our work should therefore focus on improving the SIN features first, because they do get a lot of attention from the users.

6.4 Improvements resulting from experiments

The experiments resulted in a number of observations, which we used to improve the feature extraction and presentation in a user-centric design process. As feature ranking is easier to change than feature extraction, we started by exploring different ways of ranking the extracted SIN concepts.

In the new approach, regardless of the detected or given topic, for a given video, we aim to rank higher visual concepts that have been detected with a higher confidence. We expect that these visual concepts can be used to describe a video's content. The drawback is that the general concepts like "Apartments" or "Primate", which have a higher probability in general, are ranked at top (cf. Fig. 7). To solve this problem, we can use Inverse Document Probability (IDP) to normalize the SIN score as follows

$$\mathrm{IDP}(j) = -\log(\mathrm{avg}_{k=1,\ldots,n}(\mathrm{SIN}(k, j))$$
$$\mathrm{SIN}_1(i, j) = \mathrm{SIN}(i, j) * \mathrm{IDP}(j)$$

Table 2 Comparison of naive and improved selection of feature values for the third pilot study.

Naive selection	Improved selection
Man_Made_Thing	Room
Eukaryotic_Organism	Kitchen
Body_Parts	Hand
Room	Clearing
Single_Person	Adult_Female_Human

where $\mathrm{SIN}(i, j)$ refers to the probability of the jth visual concept in the ith video of the development dataset. For further refinement, we can use average accuracy for each concept which is calculated on SIN dataset using given labels:

$$\mathrm{SIN}_2(i, j) = f(\mathrm{SIN}_1(i, j), \mathrm{AverageAccuracy}(j))$$

where $j \in \{1, 2, \ldots, 346\}$ and $f()$ refers to the product function, but other functions can also be used.

Looking at examples, SIN_1 seems to deliver better results than the other metrics, so it will be used in an updated user study. Table 2 shows an example. Besides the event-specific visual concepts, which refer to the more relevant concepts for the determined event as explained in Sect. 3.2, we also have video-specific visual concepts, which are detected with high confidence in some keyframes of the video. These can help to identify the video among all the videos in the same event.

To improve the quality of ASR transcripts, we use TF-IDF to calculate the relevant of each ASR results to the event kit, and then rank the ASR transcripts according to its relevance to the event. For ObjectBank results, we processed them in the same way as proposed for the visual concepts (see Sect. 3.2), to make them more understandable to human users and more relevant to the event. For audio concepts, we use the duration histogram of each audio concept in the video to mention that in this video we can mainly hear the sound of music, singing or noise. We use bipartite graph matching to map the 42 noisemes to the five events, so that some noisemes are more important for specific events. All the audio concepts are ranked based on their percentage in the video.

7 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 dynamic element. We presented our TOMS system, which is currently capable

of generating text-based recountings for videos belonging to one of fifteen multimedia event detection (MED) events in a database of (currently) 5,000 h. Our automatic system includes and fuses state-of-the-art audio and video features, and can be used to explain for example why a certain video is assigned to a certain topic, or how videos belonging to the same topic differ.

The aim of the presented pilot user studies is to develop a method that can be used to evaluate the performance of multimedia summarization systems directed toward indicative and informative uses. Our work provides preliminary evaluation results for text-based multimedia summarization. We propose to eventually use iterative user test, a method that can be applied to a broad set of problems, and approaches. The current demonstrator platform can be adapted for crowdsourcing experiments and demonstration purposes on the web, so our initial experimental results can and will be used as benchmarks in future work.

In this paper, we proposed and tested methods to measure the performance of a TOMS system in video selection and event selection tasks, and compared them with a human baseline. We analyzed the differences to the baseline at the level of absolute performance and observations included in the system output, and showed how these results can be used to guide future development at the "discourse" level of text generation. At a more abstract level, we then performed a study to see which of the automatically extracted features (e.g. semantic audio or video concepts, speech recognition transcripts, or objects) users tend to pay attention to, which will inform future research on which features to present, and how.

In future work, we will integrate more features to guide the recounting, further improve our system using a user-centric design incorporating the evaluation metrics of NIST's multimedia event recounting (MER) task [23], and try to scale up and automate the evaluation process, addressing both indicative and informative aspects of recounting.

Acknowledgments This work is partly supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. Support was also provided in part by the National Science Foundation (NSF) under awards IIS-0917072 and CCF-1019104, and the Gordon and Betty Moore Foundation, in the eScience project, as well as by a Faculty Research Grant from the Cisco Research Center.

References

1. Banerjee S, Rudnicky AI (2008) An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. In: Proceedings of spoken language technology (SLT). IEEE, Goa
2. Bao L, Cao J, Zhang Y, Li J, Chen MY, Hauptmann AG (2010) Explicit and implicit concept-based video retrieval with bipartite graph propagation model. In: Proceedings of the international conference on multimedia (ACM MM '10). ACM, New York
3. Bao L, Yu SI, Lan ZZ, Overwijk A, Jin Q, Langner B, Garbus M, Burger S, Metze F, Hauptmann A (2011) Informedia @ TrecVID 2011. In: Proceedings of TrecVID workshop. NIST, Gaithersburg
4. Chen FR, Withgott MM (1992) The use of emphasis to automatically summarize a spoken discourse. In: Proceedings of ICASSP. IEEE, San Francisco
5. Christel MG (2006) Evaluation and user studies with respect to video summarization and browsing. In: Proceedings of multimedia content analysis, management, and retrieval. IS&T/SPIE Symposium on Electronic Imaging, San Jose
6. Christel MG (2009) Automated metadata in multimedia information systems: creation, refinement, use in surrogates, and evaluation. Morgan and Claypool, San Rafael
7. Ding D, Metze F, Rawat S, Schulam PF, Burger S (2012) Generating natural language summaries for multimedia. In: Proceedings of 7th international natural language generation conference. ACL, Starved Rock
8. Ding D, Metze F, Rawat S, Schulam PF, Burger S, Younessian E, Bao L, Christel MG, Hauptmann A (2012) Beyond audio and video retrieval: towards multimedia summarization. In: Proceedings of ICMR. ACM, Hong Kong
9. Do Q, Roth D, Sammons M, Tu Y, Vydiswaran VV (2009) Robust, light-weight approaches to compute lexical similarity. Technical report, University of Illinois. Computer Science Research and Technical Reports
10. Dumont E, Mérialdo B (2009) Automatic evaluation method for rushes summary content. In: Proceedings of the 2009 IEEE international conference on multimedia and expo, ICME'09. IEEE Press
11. Furui S, Kikuchi T, Shinnaka Y, Hori C (2004) Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Trans Speech Audio Process 12(4):401
12. Hauptmann AG, Christel MG, Lin WH, Maher B, Yang J, Baron RV, Xiang G (2007) Clever clustering vs. simple speed-up for summarizing rushes. In: Proceedings of TRECVID video summarization workshop (TVS '07). NIST
13. Hori C, Furui S (2004) Speech summarization: an approach through word extraction and a method for evaluation. IEICE Trans Inf Syst E 87D(1):15–25
14. Jin Q, Schulam PF, Rawat S, Burger S, Ding D, Metze F (2012) Event-based video retrieval using audio. In: Proceedings of INTERSPEECH. ISCA, Portland
15. Kolb P (2009) Experiments on the difference between semantic similarity and relatedness. In: Proceedings of 17th Nordic conference on computational linguistics, NODALIDA '09. Odense, Denmark
16. Langner B, Black A (2009) Mountain: a translation-based approach to natural language generation for dialog systems. In: Proceedings of IWSDS. Irsee, Germany
17. Li H, Bao L, Gao Z, Overwijk A, Liu W, Zhang LF, Yu SI, Chen MY, Metze F, Hauptmann A (2010) Informedia @ TrecVID 2010. In: Proceedings of 2010 TrecVID Workshop. NIST, Gaithersburg
18. Li Y, Merialdo B (2010) Multi-video summarization based on avmmr. In: Proceedings of 2010 international workshop on content-based multimedia indexing, pp 1–6
19. Li Y, Merialdo B (2010) Vert: automatic evaluation of video summaries. In: Proceedings of the international conference on multimedia, MM '10. ACM, New York
20. Liu F, Liu Y (2010) Using spoken utterance compression for meeting summarization: a pilot study. In: Proceedings of spoken language technology. IEEE

21. Malkin RG (2007) Multimodal technologies for perception of humans. The CLEAR 2006 CMU acoustic environment classification system. Springer, Berlin
22. Marchionini G, Song Y, Ferrell R (2009) Multimedia surrogates for video gisting: toward combining spoken words and imagery. Inf Process Manag 45(6):616–630
23. National Institute of Science and Technology: Guidelines for TRECVID (2012) http://www-nlpir.nist.gov/projects/tv2012/tv2012.htmlmer
24. Nenkova A (2006) Summarization evaluation for text and speech: issues and approaches. In: Proceedings of INTERSPEECH. ISCA, Pittsburgh
25. Nguyen C, Niu Y, Liu F (2012) Video summagator: an interface for video summarization and navigation. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '12. ACM, New York
26. NIST Information Technology Laboratory: 2011 TRECVID Multimedia Event Detection Track (2011) http://www.nist.gov/itl/iad/mig/med11.cfm
27. Objectbank. http://vision.stanford.edu/projects/objectbank/
28. Over P (2011) Guidelines for trecvid 2011. http://www-nlpir.nist.gov/projects/tv2011/tv2011.htmlsin. NIST
29. Shen EYT, Lieberman H, Davenport G (2009) What's next? Emergent storytelling from video collection. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '09. ACM, New York
30. Snoek CG, Worring M (2008) Concept-based video retrieval. Found Trends Inf Retrieval 2(4):215–322
31. Song Y, Marchionini G, Oh CY (2010) What are the most eye-catching and ear-catching features in the video?: implications for video summarization. In: Rappa M, Jones P, Freire J, Chakrabarti S (eds) Proc. WWW. ACM, New York
32. Tan CC, Jiang YG, Ngo CW (2011) Towards textually describing complex video contents with audio-visual concept classifiers. In: Proceedings of ACM multiMedia. ACM, Scottsdale
33. Truong BT, Venkatesh S (2007) Video abstraction: a systematic review and classification. ACM Trans Multimedia Comput Commun Appl 3(1):1–37
34. Ushiku Y, Harada T, Kuniyashi Y (2011) Understanding images with natural sentences. In: Proceedings of ACM multiMedia. ACM, Scottsdale
35. Valenza R, Robinson T, Hickey M, Tucker R (1999) Summarization of spoken audio through information extraction. In: Proceedings of ESCA workshop on accessing information in spoken audio, pp 111–116
36. Wang F, Merialdo B (2009) Multi-document video summarization. In: Proceedings of the 2009 IEEE international conference on multimedia and expo, ICME'09. IEEE Press
37. Westman S (2010) Research and advanced technology for digital libraries. In: Lecture notes in computer science, vol 6273. Evaluation constructs for visual video summaries. Springer, Berlin
38. Zsombori V, Frantzis M, Guimaraes RL, Ursu MF, Cesar P, Kegel I, Craigie R, Bulterman DC (2011) Automatic generation of video narratives from shared ugc. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia, HT '11. ACM, New York