PERSONALIZED NEWS THROUGH CONTENT AUGMENTATION AND PROFILING*

Norman Haas[1], *Ruud Bolle*[1], *Nevenka Dimitrova*[2], *Angel Janevski*[2], *and John Zimmerman*[2]

[1]IBM T.J. Watson
19 Skyline Drive, Hawthorne, NY 10532

[2]Philips Research
345 Scarborough Road, Briarcliff Manor, NY 10510

ABSTRACT

This paper is concerned with the topic of personalized news assembly at the set-top box, based on augmented video. This is video complemented with additional information that is somehow relevant to the semantic video content. We touch upon the technique that is used for video augmentation, which is video subject detection followed by information searches on the subject. A focus of this paper is on subject detection implemented using traditional text analysis tools; video segmentation is based on these results and visual processing. We describe the architecture of such a system and the benefits to the consumer. Further we discuss a preliminary system that shows the viability of the concept.

1. INTRODUCTION

This collaborative work by Philips and IBM builds on a scenario for home consumer viewing, which assumes a set-top box with personal video recorder (PVR) capability, also enabled with Internet access of modest bandwidth, such as dial-up, DSL, or cable back-channel. The box is assumed to hold a user profile, describing its geographical location and the viewing preferences and interests of its user. A preliminary version of such a system was demonstrated at NAB 2002.

The goal of the set-top box is to automatically monitor one or more television broadcasts (cable and/or terrestrial) and multiple Internet information sources, and to always have an optimal news presentation ready for the viewer, whenever the user should desire to see it. Monitoring multiple channels is possible even with a set-top box with a single tuner; it can use EPG (Electronic Program Guide) information, specifically, program type information, to sequentially tune to news programs on various channels. The set-top box will be able to perform many intelligent functions, but in this paper, we limit the discussion to its handling of the personalized news application.

2. PERSONAL NEWS APPLICATION

The personal news application should differ from standard news broadcasts in its level of user accommodation, as well as in its content. Newsgathering from both TV and the Web should be seamlessly integrated; users should not have to use two separate devices/modalities to get all their information. And there should be near-zero latency; users do not want to have to wait to get the weather or traffic report. Further, a successful personal news application ought to offer immediate access to the freshest information, and it should be categorized into a few catch-all categories, such as weather, traffic, sports, financial, local events, and general headlines. Also, since viewers often watch the news while doing other tasks -- in the morning, for instance, they may be preparing their breakfast and conducting other maintenance tasks -- a hands-free playback mode is important. And where interaction is required, a TV remote control-like interface is preferable to keyboard/mouse.

The style of interaction with such a system would be about halfway between pure television and pure web browser: large-format or full-screen video images, yet with more text and more interaction than television alone, but less than full-blown web-surfing. The user's body position while using this system would be neither "lean-forward" nor "lean-back," but "lean-natural."

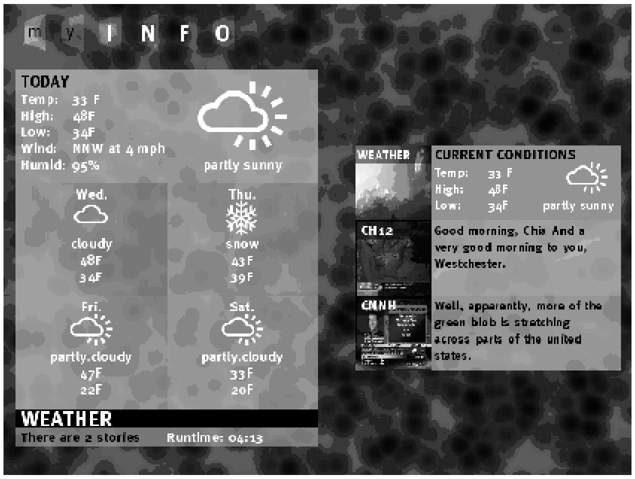
We have developed a system in response to the above criteria, which allows users to select any one of the six "content zones" (categories) identified above, and immediately see a list of all the TV stories related to that zone plus personalized data relating to the zone that has been extracted from the Internet.

Figure 1 shows the display that might be presented for the Weather zone. Today's weather and a five-day forecast extracted from the web are on the left side of the screen, while the right shows the list of all the weather-related stories, i.e., both Web and TV. Up/down buttons on the remote scroll this list, and the story which moves into the highlighted position, if it is a TV story, is played on the left, or, at the user's request, full-screen. For hands-free playback, all stories for the currently active zone may be played, or all news stories in all categories can be played, in a prioritized order (explained later).

3. THE CONTENT AUGMENTATION SYSTEM

To support this vision of a personalized news service, a content augmentation system with an architecture such as is shown in Figure 2, is desirable.

* Many other researchers contributed greatly to this project: Lalitha Agnihotri, Dongge Li, Thomas McGee, Radu Jasinschi, Jonathan Connell, and Senem Velipasaran.



TODAY
Temp: 33 F
High: 48F
Low: 34F
Wind: NNW at 4 mph
Humid: 95%
partly sunny

Wed.
cloudy
48F
34F

Thu.
snow
43F
39F

Fri.
partly.cloudy
47F
22F

Sat.
partly.cloudy
33F
20F

WEATHER
There are 2 stories Runtime: 04:13

CURRENT CONDITIONS
Temp: 33 F
High: 48F
Low: 34F
partly sunny

Good morning, Chis And a very good morning to you, Westchester.

Well, apparently, more of the green blob is stretching across parts of the united states.

Figure 1: Weather screen

At the point of broadcast, ordinary program content (MPEG-2 encoded, perhaps with some metadata already inserted, say, in MPEG-7 format), is analyzed and augmented with information that concerns most or all of the targeted viewer groups. This includes labeling to describe program segments, and data acquired from interactions with various Internet sources, triggered by the program content. All augmentation is embedded as program metadata in the outgoing bitstream, which is received by the set-top box at the end of the delivery chain.

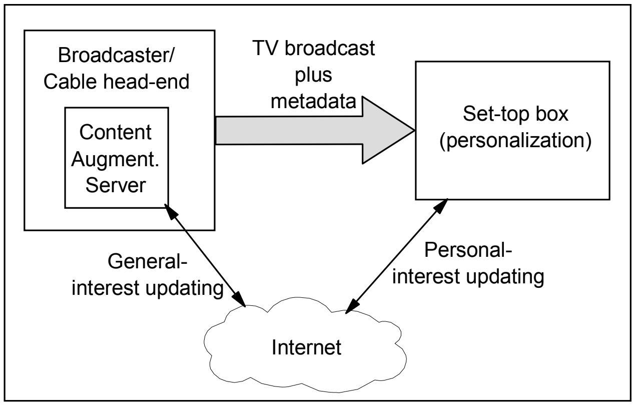


Figure 2: The content augmentation paradigm

At the set-top box, received content is stored, the segments of which then become available for time-delayed viewing, in random-access order. Received metadata is used to prioritize program segments for playback according to their importance and relevance with respect to the user profile, and additional web interactions for content augmentation is performed, to further personalize the news (in particular, by acquiring locality-specific information). These interactions are periodically re-executed, to refresh the supplemental information, for as long as the program segments are retained.

Since the computation-intensive processing is performed in the content augmentation server, the set-top boxes do not require expensive processors. And since the C. A. server also performs the Web interactions to acquire the supplemental data of general interest, the Internet is not loaded down with redundant traffic – the set-top boxes only query for information specific to their individual users. Lastly, by keeping personalization in the set-top box, users' privacy remains intact.

Figure 3 (next page) shows the architecture of the content-processing chain of the demonstrator shown at NAB. Details will be discussed in Section 4.

4. PROGRAM CONTENT ANALYSIS

Content analysis is essential for developing the necessary high-level information from the image pixels and waveform samples of the source program. For the personal news application, there are basically two sets of required tasks:

1. Segmentation, classification and summarization of stories, and, as a subtask of this, the detection of embedded non-news material, such as station identification and commercials. This is performed entirely in the broadcaster's content augmentation server, in Figure 2.
2. Determination of story relevance and priority, with respect to location and term (category; name) information in the user personal profile. This is performed at the individual set-top box, but it involves matching with keywords (names) in the story, which are identified by the content augmentation server.

News story segmentation is a difficult problem, because it involves a somewhat subjective judgement call. Slaney [8] has nicely addressed this problem from a scale-space point of view. In this work, however, we prefer a flat segmentation model.

Based on advances in image, audio and text analysis techniques, we believe fully automatic analysis will become possible, by integrating many technologies. The techniques that we have worked with in the demonstrator are:

- closed-caption extraction,
- speech transcription,
- text summarization and categorization,
- anchorperson detection,
- simple image features (indoor / outdoor), and
- multimedia summarization.

4.1 Audio and transcribed text features

We argue that, for news, the bulk of the useful audio information resides in the spoken words themselves, rather than in how they were spoken, or in the non-verbal sounds. For example, Smith and Kanade is a good indexing system for videos (not limited to news) which uses audio, video and text modalities [9]; we note that it puts most of the indexing burden on the text modality.

So, in particular, the problem of story segmentation of a news program can be attacked by reducing it to the problem of transcribing the program's dialogue, segmenting the transcript, and then perhaps doing something about the temporal

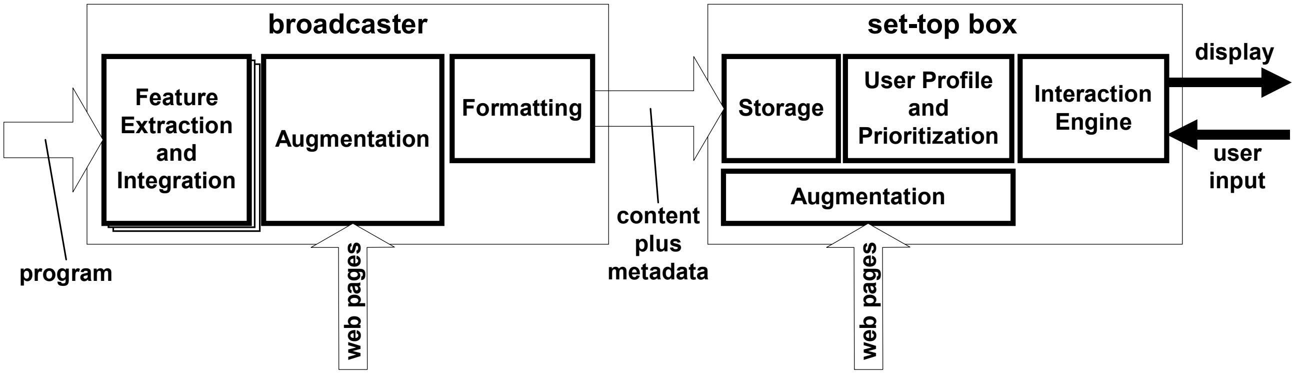


Figure 3: Content augmentation system architecture

misalignment between the timings of the transcribed words and their spoken equivalents.

Speech transcription per se is a relatively mature technology [10]. General audio track transcription is harder, with all sorts of possible ambient noises (e.g., music). Accuracy is steadily improving, but closed-caption text (i.e., human-generated transcription) is still to be preferred whenever it is available. A comprehensive and excellent overview of automated speech recognition for indexing applications is presented in Coden et al. [3]. We note that all of the indexing, summarization, segmentation, and categorization of video, and the subject detection technologies based on automated speech recognition described in [3], use little or no visual information.

Once the transcript is in hand, ample guidance on what to do next is available. There is a considerable body of work on "discourse segmentation" in the computational linguistics literature. ("Discourse" refers to both printed and spoken lengthy utterances).

One popular recurring idea is to partition the discourse into fragments, and to measure the similarity of one fragment to another, using the cosine metric, which is the dot product of the word occurrence frequencies (after morphological analysis, to reduce words to their base form, and sometimes also using a thesaurus or semantic net to allow different words having with similar meanings to match). Other systems take into account occurrence of multimedia features, such as cue phrases and prosodic features (Litman and Passoneau).

For segmentation for our demonstrator, we kept things simple and relied on the delimiters inserted in the closed captioning ('>>>').

Once segmented, each story must be categorized, summarized and prioritized. For the first two of these, we used a commercial product [5] to select, from each document, the *n* sentences which in its opinion are the best summarization of that document (we used $n = 1$). We also used it to categorize documents according to a taxonomy that we had previously defined by training on a corpus of hand-classified documents; we trained it with the six content zones of Section 2, plus "commercial", "opening", "closing", and "miscellaneous".

We also used the product [5] to detect and distinguish names of places, persons and entities from common words. We used the identified names to determine a story relevance score, by matching names in each story against those in the user profile. To obtain a measure of the story's overall priority for playback to the user, we combined this score with a measure of the "objective" importance of the story. The objective importance is based on:

- The position of the story within the news program
- The amount of time devoted to the story.
- The number of other news sources carrying the story.

(The above assumes proper capitalization of the text. Capitalization may be present in automatically extracted transcripts, but closed-captioned text is generally monocase. Brown and Coden [1] describe a system for re-capitalizing monocase texts.)

4.2 Visual features

Visual features also contribute to story segmentation, summarization, and classification, and we exploited this:

Scene color analysis [4] contributes to scene classification and visual reportage summarization (key frame detection). We make heavy use of it.

Face detection has been extensively worked on as a computer vision topic [2]. We have an anchorperson detection module using techniques like face detection. This is an important contributor to multi-modal segmentation, since stories often begin and/or end with in-studio (e.g., anchor-present) shots, rather than during "reportage" segments. Further, one aspect of story summarization is selection of a key frame to be used as a graphic on the zone-summary screen. It is desireable to chose this frame from the reportage, rather than from a shot of the anchorperson, since the latter will look the same for all stories.

Indoor/outdoor, and color graphic detector [7]. The former contributes to detecting anchorperson sequences, for segmentation; the latter helps avoid bad key frame selection during summarization.

5. WEB INFORMATION EXTRACTION

At both the broadcasting site and the set-top box, augmentation from Internet web sites is performed. By using Web Information Extraction (WebIE), a rule-based system, structured information is automatically derived from ordinary worldwide web (WWW) documents [6]. In a goal-driven manner, WebIE rules are applied, creating tasks to (i) retrieve a document, (ii) segment it at the level of tags and keywords, (iii) do some limited parsing, (iv) extract raw desired information, and (v) perform data cleanup and regularization. The rules allow for each task to be customized for the specific type of data to be retrieved.

In the set-top box, "WebIE" is used to acquire weather conditions and forecasts; traffic routes and hot spot reports; index, stock, and fund quotes; sports results and event schedules; and local events announcements. The source documents, published on the web by enterprises not necessarily associated with the broadcasters, need not be structured (beyond being written in HTML), yet they are narrowly defined, both through their locations (URLs) and the structural organization of their content. For instance, localization for weather may be expressed by occurrence of zip code as part of the weather pages' URLs; Laser WebIE can synthesize the appropriate string and get the exact page it needs, without searching; hence the name "Laser," to distinguish it from "Diffusion WebIE," which crawls (explores) the WWW.

6. DISCUSSION

The concept of this system was tested with a focus group, and was well-received. A preliminary demonstration of concept implementation was exhibited at NAB this past April.

Results with the automatic enabling technologies are so far, good but mixed. One challenging characteristic of the news application is its fluidity: over time, and across stations, many aspects of the input data change: directorial styles of presentation, relating to screen formats (CNN's switch to quarter-screen images of the anchor), view and number of anchors (CNN shows one, full-face, whereas Channel 12 NY shows views from the side as well as front, sometimes with anchor plus weather or traffic person), narration style (CNN stories always start and end with camera on the anchor, whereas Channel 12's do not), etc.

A conclusion of the work is that in order to make advances in the area of automated program analysis, more attention must be paid to text. Text is a very important - in fact, the dominant - modality. Program segmentation and genre can be determined practically from text alone. Once they are determined, they can be used to select the category-specific audio and video feature detectors appropriate to be applied to the content in question. Even if segmentation and genre are not determined from internal content features, but are determined by reading the Electronic Program Guide (EPG), that is another form of control by text.

7. REFERENCES

[1] E.W. Brown and A. R. Coden, "Capitalization Recovery for Text," A. R. Coden, E.W. Brown, and S. Srinivasan (eds.), *Information Retrieval Techniques for Speech Applications*, Springer, New York, pp. 11-22, 2002.

[2] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," Proc. IEEE, Vol. 83, pp. 705-740, May 1995.

[3] A. R. Coden, E.W. Brown, "Speech Transcript Analysis for Automatic Search," IBM Research Tech. Rep. RC-21838, Sept., 2000.

[4] N. Dimitrova, L. Agnihotri, and G. Wei, "Video classification based on HMM using text and faces," Proc. European Conference on Signal Processing, Finland, 2000.

[5] IBM Intelligent Miner for Text™.

[6] A. Janevski and N. Dimitrova, "Web Information Extraction for Content Augmentation," To appear in Proc. IEEE International Conference on Multimedia and Expo, Switzerland, Aug., 2002.

[7] M. R. Naphade, I. Kozintsev, and T. S. Huang, "A Factor Graph Framework for Semantic Video Indexing", IEEE Transactions on Circuits and Systems for Video Technology, Vol.12, No.1, pp. 40-52, Jan. 2002.

[8] M. Slaney, D. Ponceleon, and J. Kaufman, "Multimedia Edges: Finding Hierarchy in all Dimensions," Proc. 9th ACM International Conference on Multimedia, Ontario, Sept./Oct., 2001.

[9] M.A. Smith and T. Kanade, "Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques," Proc. DARPA Image Understanding Workshop, New Orleans, May, 1997, pp. 357-366.

[10] IBM ViaVoice Millenium Pro. 2000.