

# Semi-Automatic News Analysis, Indexing and Classification System based on Topics Preselection<sup>^</sup>

Alan Hanjalic, Reginald L. Lagendijk, Jan Biemond

Delft University of Technology  
Faculty of Information Technology and Systems  
Information and Communication Theory Group,  
P.O. Box 5031, 2600 GA Delft, The Netherlands  
{alan,inald,biemond}@it.et.tudelft.nl

## ABSTRACT

In this paper we present the concept of an efficient semi-automatic system for analysis, classification and indexing of TV news program material and show the feasibility of its practical realization. The only input into the system, other than the news program itself, are the spoken words, serving as keys for topic prespecification. The chosen topics express user's current professional or private interests and are used for filtering the news material correspondingly. After the basic analysis steps on a news program stream, including the processes of shot change detection and key frame extraction, the system automatically represents the news program as a series of longer higher-level segments. Each of them contains one or more video shots and belongs to one of the coarse categories such as anchorperson (news reader) shots, news shot series, the starting and ending program sequence. The segmentation procedure is performed on the video component of the news program stream and the results are used to define the corresponding segments in the news audio stream. In the next step, the system uses the prespecified audio keys to index the segments and group them into reports, being the actual retrieval units. This step is performed on the segmented news audio stream by applying the wordspotting procedure to each segment. As a result, all the reports on prespecified topics are easily reachable for efficient retrieval.

**Keywords:** TV news retrieval, video retrieval tools, video indexing, video databases, video content analysis, wordspotting

## 1. INTRODUCTION

It has been widely recognized that the TV news programs are highly interesting "storing objects" in emerging large-scale digital video databases. The main reason lies undoubtedly in their information content, which may be useful for applications in many professional areas as well as for user-private needs. One could think of building up large information archives, containing all available sorts of informative programs, e.g. news, documentaries, TV-debates, political or social discussions, reportages, etc. In such archives, news are at least as important as all other mentioned program types, since they concisely cover huge amounts of topics related to society, daily politics, sports, business, etc. The importance of news programs may even be larger, since not all daily events get a thorough coverage through e.g. a dedicated documentary. Collecting news over a longer time period from different broadcasters can therefore provide a solid top level for an information collection, whereby other informative programs on certain topics, if any, are linked to relevant news reports and serve as lower-level (more detailed) information sources.

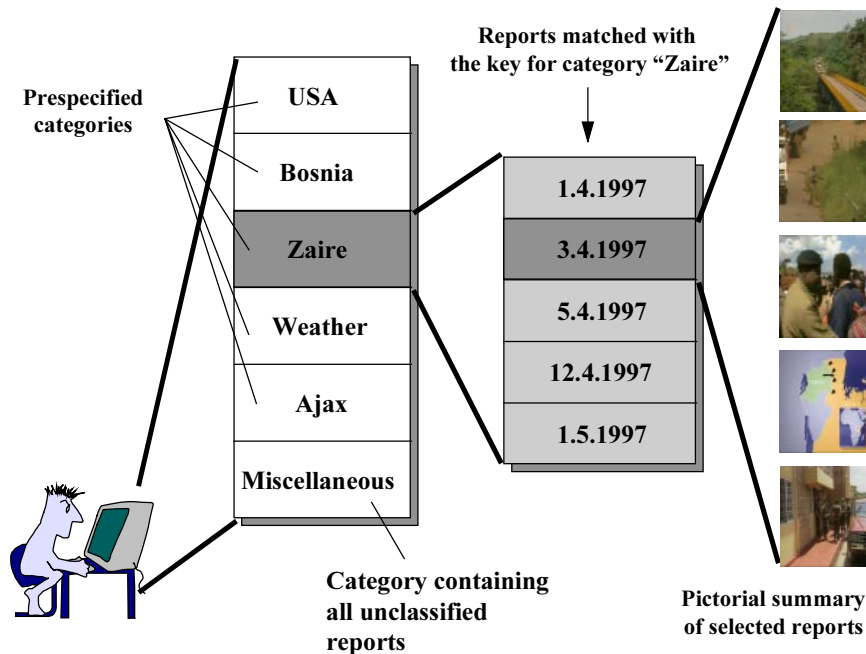
If concentrating on news programs, it is highly important for the practical usability of a large-scale news archive that a report on any given topic can be retrieved quickly and reliably. Also different retrieval wishes have to be taken into consideration, like e.g. "find me a business report in a CNN news program from 2.4.1997", or "give me everything what is available on car races". It is not difficult to think of an ideally automated news analysis, indexing and classification system, which automatically segments a news program into different reports, recognizes their topics and classifies them correspondingly. In that case, after obtaining the list of extracted topics from the stored news material, the user can immediately start the interaction with the system by browsing through the organization tree, selecting the topics of interest

---

<sup>^</sup> This work was supported in part by the EU ACTS program under the contract AC018: *SMASH* (Storage for Multimedia Applications Systems in the Home)

and retrieve any of the news reports on those topics. He can also perform queries by example, by submitting the topic of interest to the system and retrieving relevant reports if the requested topic was found within the stored material. Several approaches to news program analysis, indexing and retrieval aiming at such ideal solution can be found in recent literature [1, 2, 4, 5, 8, 12, 15, 19]. Unfortunately, a full automation of the presented system is possible only theoretically, since it is beyond current technical state-of-the-art. The actual problematic step in this idealized system is the automated topic recognition for an arbitrary news segment.

However, a semi-automatic option, where the topic recognition is driven by *topic-specific keys* given *a priori* by the user, appears to be technically feasible. Such keys are the only input into the system besides the actual news program. The list of topics covered by the keys can differ for various professional or private areas of interest. The output of such a system can allow for the close-to-ideal retrieval scenario, as illustrated in Figure 1. There, the first interaction level shows the list of preselected categories (topics), in this case five of them (USA, Bosnia, Zaire, Weather and Ajax), which are linked to all corresponding reports found in the entire stored news material. As an example in Figure 1, five reports have been found which matched with given key(s) for the topic “Zaire”. Each of these reports can then be retrieved, after looking at its preliminary representation given by e.g. its pictorial summary using a number of key frames. All the remaining news material, which could not be matched with any of the prespecified topics, is linked to the category “miscellaneous”. The user can easily browse through the obtained structure, or perform a query, by giving in a specific key, which is now matched with keys already present in the system. If a match is found with any of categories, the corresponding set of reports linked to this category can be retrieved.



**Figure 1:** Retrieval of news reports on specific, prespecified topics

The purpose of this paper is to show the technical feasibility of the presented semi-automatic system for analysis, indexing and classification of a news material and to propose a suitable realization scheme. The basic retrieval unit in the scheme of Figure 1 is a single news report on a prespecified topic. Therefore, the major task for the proposed analysis and indexing system is to locate all the reports corresponding to given topics and to find their exact boundaries. We approach this task in two steps. In the first step, described in Section 2 and performed on the news video stream, potential report boundaries are found by classifying all video shots of the news program in longer segments, which can be assumed to be global report components. In the second step, described in Section 3 and performed on the news audio stream, segment topics are identified by applying the prespecified keys. The related segments are merged into reports and the remaining segment boundaries after the merging procedure are the precise report boundaries. In Section 4, a system realization scheme is presented, accompanied by the experimental evaluation of the news segmentation, segment identification and report forming procedure. A discussion about the material presented in this paper can be found in Section 5.

## 2. HIGHER-LEVEL NEWS PROGRAM SEGMENTATION

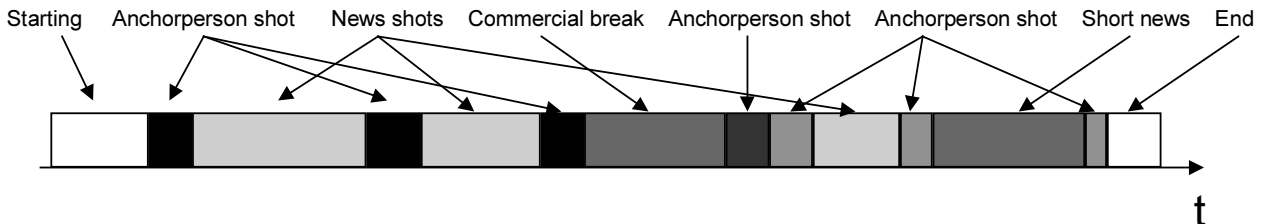
Assuming that the basic video analysis procedures on a news video stream, such as shot change detection and key frame extraction, are completed, we can generally classify all the video shots of a news program into segments, which belong to the following coarse categories:

- An anchorperson shot (showing one or two news readers)
- A news shot series (a series of shots taken by a reporter on a site outside the studio)
- The starting or ending news sequence
- A commercial break (not always present)

Since the report structure can generally be defined as a combination of anchorperson shots and news shot series, only the segments belonging to these two categories are taken into account by the report forming procedure.

A realistic structure of a news program in view of the four different categories is illustrated in Figure 2 [19]. According to the structure, it is obviously possible to find the boundaries of all news segments by detecting anchorperson shots and commercial breaks. The starting and ending news sequences can then be determined directly by the first and the last anchorperson shot, while a segment lying between two anchorperson shots is a news shot series, if not previously identified as a commercial break. While some work on detection of commercials in TV programs has already been done [11], we concentrate here on the problem of automatically detecting *anchorperson shots* in an arbitrary news program and propose a new approach for performing this operation. Without losing generality, we assume from now on that no commercial breaks are present in the news programs.

Compared to already existing anchorperson shot detection methods (e.g. [2, 8, 15]), we believe to achieve an increase in detection robustness due to the usage of a sequence-own video shot as the template for detecting all anchorperson shots of that sequence. The template itself is also found in a robust, threshold-free way. When defining our detection method, we make use of the news program structure as shown in Figure 2, and take into account different types of anchorperson shots, characterized by the number of news readers, camera position, etc.



**Figure 2:** Assumed realistic structure of a news program. Different gray values of blocks belonging to anchorperson shots indicate their different types

### 2.1 Assumptions and definitions

We base our anchorperson shot detection approach on the assumption that an anchorperson shot is the only type of video shots in a news program that has multiple matches of most of its visual content along the entire news program. Other (news) shots may match well only in their closest neighborhood (e.g. within a single report) where they can eventually find enough similar visual features. Such an assumption is realistic due to specific visual characteristics of anchorperson shots and their regular appearance along a news sequence. We also assume that the first anchorperson shot  $k_{ap}$  in a news program containing  $S$  video shots certainly appears within the interval  $[1, N]$ , where  $N < S$  is assumed to be around 5. In order to make the detection as robust as possible, we took into account different types of anchorperson shots, also including non-stationary ones. We introduce now the following definition:

*Anchorperson (news reader) shots are visually characterized by studio background and by one or two news readers appearing separately or together, also with some possible variations of a camera angle and changes in news icons appearing in a screen corner. These shots can be static or dynamic (containing some camera operations like zooming or panning) and generally contain a certain (high) percentage of same or similar visual features.*

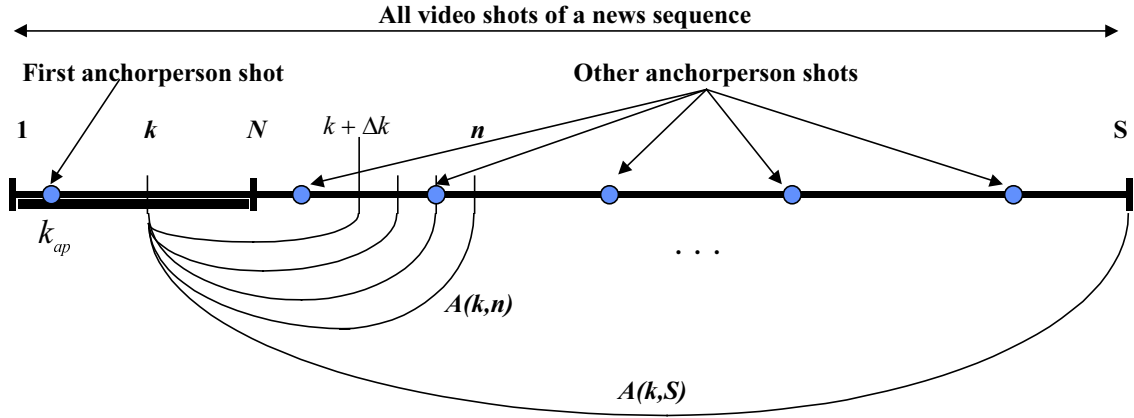
During the detection procedure we compare video shots based on their visual abstracts. Hereby, we assume that, prior to the anchorperson shot detection procedure, a news sequence has already been segmented into video shots, and that each shot is represented by a visual abstract consisting of a limited number of *key frames*.

The proposed anchorperson shot detection approach consists of two steps, described in Sections 2.2 and 2.3 respectively:

- A threshold-free procedure of finding the sequence-specific template for anchorperson shots,
- Using the template to detect all anchorperson shots in a sequence by applying adaptive thresholding.

## 2.2 Finding a template

Figure 2 shows the possibility of different types of anchorperson shots appearing together with news shots series and commercial breaks. Based on assumptions from Section 2.1 we match each shot  $k \in [1, N]$ ,  $N < S$ , with all other news shots  $n \in [k + \Delta k, S]$ , as shown in Figure 3. In this way, a set of dissimilarity values  $\{A(k, k + \Delta k), \dots, A(k, S)\}$  is obtained for each shot  $k$ . The dissimilarity measure used here to compute values  $A(k, n)$  compares two shots on basis of their abstracts (key frames) and will be defined in Section 2.4. The “security” interval  $[k, k + \Delta k]$  serves to avoid a possible good match of a news shot in its neighborhood and, consequently, to separate the shot  $k_{ap}$  even stronger from the rest.



**Figure 3:** Obtaining a dissimilarity values set for the shot  $k$

For each shot  $k \in [1, N]$  we now take the  $P$  best matches (lowest values) out of its set of dissimilarities and average them to compute its overall matching value. The shot with the lowest overall matching value is assumed to be an anchorperson shot, and is used as the template for finding all other anchorperson shots of a news sequence.

## 2.3 Template matching

After the template has been found, all shots of a sequence are checked for being anchorperson shots by using the same inter-shot dissimilarity metric  $A(k, n)$ . Low dissimilarity values will be obtained when the template is matched with another anchorperson shot. For each shot  $k$  of a sequence we now define its similarity with the template shot as

$$s(k) = \frac{1}{A(t, k)} \quad (1)$$

whereby  $A(t, k)$  is the dissimilarity between the template  $t$  and the shot  $k$ . In order to perform the detection of anchorperson shots automatically, we use the adaptive threshold  $M(k)$ , defined as follows:

$$M(k) = \frac{w}{N_k + 1} \left( \sum_{i=1}^{N_k} s(k-i) + s_0 \right) \quad (2)$$

Here  $w$  is a fixed parameter whose value is not critical in a wide range of values. The parameter  $N_k$  denotes the number of shots until  $k$  and since the last detected anchorperson shot. It also uses the similarity value  $s_0$  of the last shot of the last

news report as a bias. For each shot  $k$ , a value  $s(k)$  is available as well as the threshold value  $M(k)$ . An anchorperson shot is detected when  $s(k) > M(k)$ .

## 2.4 The dissimilarity metric

We now define the dissimilarity metric  $A(k, n)$  with the intention to reliably distinguish all anchorperson shots from the rest, but at the same time to allow for slight differences among different types of anchorperson shots. As already mentioned in Section 2.3, the metric  $A(k, n)$  measures the dissimilarity between shots  $k$  and  $n$  by using their visual abstracts, which consist of key frames. We represented each shot of a news sequence by two key frames, one standing close to the beginning and one close to the end of the shot.

For each shot, both key frames are merged together into the *shot image*, which is then divided into blocks of  $M1 \times M2$  pixels. Each block is now a simple representation of one element of the shot's visual content. Since we cannot expect an exact shot-to-shot matching in most cases, and the influence of unimportant shot content details should be as small as possible, we choose to use only those features that describe the  $M1 \times M2$  elements *globally*. In this paper we use only the average color in the  $L^*u^*v^*$  uniform color space as a block's feature.

For each shot pair  $(k, n)$  we now would like to find the mapping between the blocks  $b_k$  and  $b_n$ , each being an  $M1 \times M2$  block from the shot image  $k$  and  $n$ , respectively, such that

- each block  $b_k$  in a key frame of shot image  $k$  has a unique correspondence with a block  $b_n$  in shot image  $n$ . If a block  $b_n$  has already been assigned to a block  $b_k$  from a key frame belonging to shot image  $k$ , we do not allow it to be used for matching of any other block from that key frame. All blocks  $b_n$  are available only when a new key frame of shot  $k$  is to be matched. Figure 4 illustrates this in more details.
- the average distance in the  $L^*u^*v^*$  color space between corresponding blocks from the two shot images is minimized:

$$\min_{\text{all possible block combinations}} \sum_{\text{all blocks}} d(b_k, b_n) \quad (3)$$

with

$$d(b_k, b_n) = \sqrt{(L^*(b_k) - L^*(b_n))^2 + (u^*(b_k) - u^*(b_n))^2 + (v^*(b_k) - v^*(b_n))^2} \quad (4)$$

and where all possible block combinations are given by the first item.

Unfortunately this is a problem of high combinatorial complexity. We therefore use a suboptimal approach to optimize (3). The blocks  $b_k$  from a key frame of a shot  $k$  are matched unconstrained with blocks in shot image  $n$  starting with the top-left block in that key frame, and subsequently line-fashioned scanning to its bottom-right block. A block  $b_n$  that has been assigned to a block  $b_k$  is no longer available for assignment until the end of the scanning path. For each block  $b_k$  the obtained match yields a minimal distance value  $d_1(b_k)$ . Then, this procedure is repeated for the same key frame in opposite scanning fashion, i.e. from bottom-right to top-left, yielding a difference mapping for the blocks  $b_k$  and a new minimal distance value for each block, denoted by  $d_2(b_k)$ . On the basis of these two different mappings for each key frame from shot  $k$  and corresponding minimal distance values  $d_1(b_k)$  and  $d_2(b_k)$  per block, the final correspondence and actual minimal distance  $d_m(b_k)$  per block is constructed as follows:

$$\bullet \quad d_m(b_k) = d_1(b_k), \text{ if } d_1(b_k) \leq d_2(b_k) \quad (5a)$$

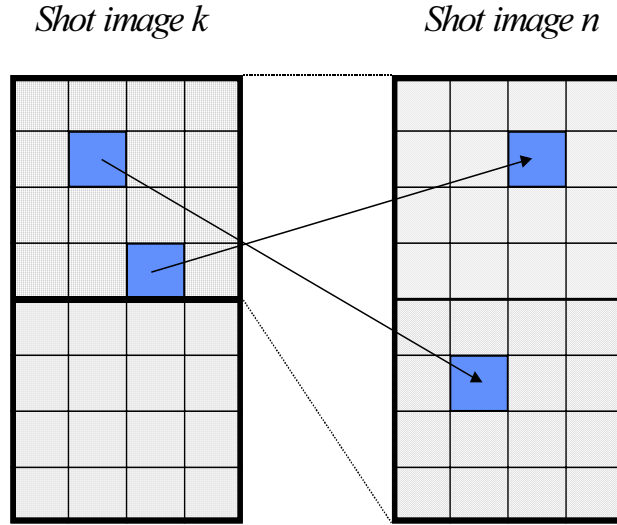
$$\bullet \quad d_m(b_k) = d_1(b_k), \text{ if } d_1(b_k) < d_2(b_k) \text{ and } d_1(b_k) \text{ is the lowest distance value measured on the assigned block in the shot image } n \text{ (one block in shot image } n \text{ can be assigned to two different blocks in a key frame from } k \text{: one time in each scanning direction)} \quad (5b)$$

$$d_m(b_k) = \infty, \text{ otherwise.} \quad (5c)$$

$$\bullet \quad d_m(b_k) = d_2(b_k), \text{ if } d_2(b_k) < d_1(b_k) \text{ and } d_2(b_k) \text{ is the lowest distance value measured on the assigned block in the shot image } n \quad (5d)$$

$$d_m(b_k) = \infty, \text{ otherwise.} \quad (5e)$$

where  $\infty$  stands for a fairly large value, indicating that no acceptable best match for a block  $b_k$  could be found. The entire described procedure is repeated for another key frame of a shot  $k$ , so that finally one value  $d_m(b_k)$  is assigned to each block of a shot image  $k$ .



**Figure 4:** Comparison of a shot  $k$  with a shot  $n$  by matching  $M1 \times M2$  blocks from each key frame of shot  $k$  with all blocks in the shot image  $n$ .

Finally the average of the distances  $d_m(b_k)$  of the best  $C$  matching blocks in the shot image  $k$  is computed as the final inter-shot dissimilarity value:

$$A(k, n) = \frac{1}{C} \sum_{b=1}^C d_m(b_k) \quad (6)$$

The reason for taking only the  $C$  best matching blocks is that two shots should be compared only on a global level, allowing for differences among different types of anchorperson shots.

### 3. REVEALING THE REPORT STRUCTURE

Once the program segments corresponding to anchorperson shots and news shot series are isolated, the compliance with the prespecified topic-specific keys has to be checked for each of them. If such compliance is found, the segment is identified and merged with other related segments into a report. The formed report remains linked to the corresponding topic for the retrieval procedure.

The topic-specific keys have to be selected in such a way, that there is a high probability for matching between the key and all the corresponding reports. In this sense, visual keys are not very suitable since it is not possible to say which visual feature is most characteristic for a certain topic, and therefore most likely to appear in a report. Two reports on the same topic can have totally different visual content (e.g. war report from a site, and interview with a UN official about that war). For this reason, the user cannot know when predefining a topic, what visual features will appear in the relevant report. There is also the possibility to use textual keys and match them with captions appearing on the top or bottom of a screen during the report. However, since the appearance of such captions is dependent on a specific news program provider, this doesn't seem to be a highly reliable option. We found the audio (speech) keys in form of spoken words being the most reliable for topic recognition. The reason is quite clear: independent of what visual content will appear in a report over Bosnia or if the report is labeled by a caption or not, the word "Bosnia" will be mentioned at least once during the report.

Taking this as the most acceptable solution for topic recognition, we can now define the user-input into the system consisting of a total number of topics and a set of well-chosen characteristic words describing the topics of interest. Tools required for matching the spoken words with the audio stream of the news program already exist and can be referred to

under the name *wordspotting algorithms* [3, 6, 9, 10, 13, 14, 17]. Generally speaking, these algorithms are capable of recognizing the appearance of user-specified key words in a continuous speech. Increasing quality of wordspotting algorithms can be noticed when investigating the research in this area over the past years. This quality is especially related to two major issues, which are also crucial for the news indexing application presented in this paper:

- Freedom of key word choice
- Speaker independence

Related to this is the problem of different languages, dialects, accents, etc. which also has to be solved, in order to successfully employ the wordspotting in practice. The speaker independence is highly important for our system since a user-spoken key word has to be matched with the same word, spoken by the news reader. Two utterances of the same word have different time-structures due to differences in pronunciation, resulting in some parts of the word being longer, some being shorter and some simply different [3].

We assume now to have available a robust wordspotting algorithm,  $R$  news segments found by the segmentation procedure in Section 2, and  $T$  prespecified topics, each described by a number of key words. The first step in the segment identification (indexing) procedure is performed by matching all segments with all key word sets. We define the value  $W(i,j)$  as the distance between a segment  $j$  and a key word set belonging to the topic  $i$ , being a function of single distances of all key words belonging to that set. The results of such a matching procedure can be presented in general as shown in Table 1.

	Segment 1	Segment 2	...	Segment R
Topic 1	$W(1,1)$	$W(1,2)$	...	$W(1,R)$
Topic 2	$W(2,1)$	$W(2,2)$	...	$W(2,R)$
Topic 3	$W(3,1)$	$W(3,2)$	...	$W(3,R)$
...	...	...	...	...
Topic T	$W(T,1)$	$W(T,2)$	...	$W(T,R)$

**Table 1:** Matching results between  $R$  news segments and  $T$  prespecified topics

Since a value  $W(i,j)$  is defined as a distance, the matching quality between a segment and a key word set increases as  $W(i,j)$  decreases. In order to make the topic recognition more robust and reduce the number of false alarms (e.g. the name of the soccer club “Ajax” just mentioned in a report over another soccer club “Bayern”), several key words should be specified to describe one single topic. In that case, a good match within a news segment needs to be found for more than one key word in order to receive a topic index. Having such requirement, a suitable way of obtaining the  $W(i,j)$  values can be defined as averaging of matches of all key words belonging to that set. On the other hand, an “over-specification” of a topic by too many key words also seems to be unsuitable, since the probability grows that more words will not find their match with the corresponding news segment. Consequently, the overall matching quality between the topic and the “right” segment declines since the average matching value  $W(i,j)$  gets worse. We found in our tests two to three keywords to be optimal for topic specification.

Based on  $W(i,j)$  values in Table 1, we now discuss the problems of segment identification (recognizing its topic), and report forming by merging related segments. Generally speaking, one topic can be assigned to several different news segments. In that case, a number of lowest  $W(i,j)$  values has to be recognized and split from the rest. Such a procedure is possible only after defining a suitable threshold value. However, being aware of difficulties in finding a threshold and evaluating its robustness in a general case, we propose here a threshold-free method for segment identification, which requires only the definition of the maximum number of segments belonging to a single report. We denote this parameter as  $G$ . We choose the topic  $q$  and investigate the  $W(q,j)$  values,  $j=1..R$ , searching for the minimum. The segment  $s$  where the minimum is found serves as the initial report on topic  $q$ , denoted as  $[s]$ . This can be formulated mathematically as

$$W(q, s) = \min_{e=1..R} W(q, e) \Rightarrow [s] \cong q \quad (7a)$$

In order to check if the report on topic  $q$  is limited only on segment  $s$ , we investigate  $G$  neighboring segments on each side of  $s$ , compare them to all topics and require that their best match is found by the topic  $q$ . Mathematically,

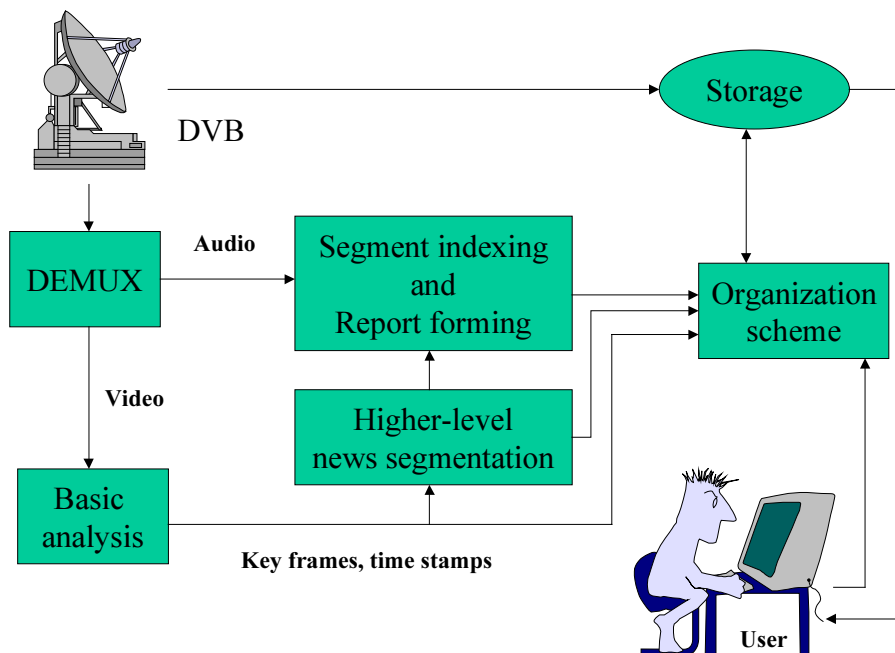
$$W(q, s-g) = \min_{e=1..T} W(e, s-g) \quad (7b)$$

The requirement (7b) does not have to be fulfilled for all values  $g=-G..G$ , having in view the fact that the given key word set does not have to be present in every report segment. If  $s-g1$  and  $s-g2$  are the farthest segments in respect to the segment  $s$  on its both sides, for which the condition (7b) is fulfilled, the report on topic  $q$  is defined as the set of segments within the interval  $[s-g1, s-g2]$ .

The described procedure is repeated for all prespecified topics and leaves the possibility open that a news segment is assigned several different topics and that it appears in different reports. This is often the case in multi-topic anchorperson shots and if the news program contains the “short news” segment, i.e. a concatenation of news shot series on different topics without anchorperson shots in-between. In case that a topic is prespecified, which is not covered by the stored news material, the proposed procedure will result in an unpredictable report structure, which can be interpreted as “the best” that the system could find for the specified topic.

#### 4. SYSTEM REALIZATION AND EXPERIMENTAL EVALUATION

In view of practical applicability of our system and our activities within the European Research Projects SMASH [16] and STORIT [17], we see the proposed news analysis, indexing and classification system as a DVB [7] application. Consequently, an MPEG program stream can be assumed as system input. After demultiplexing operation, the video stream is processed first by performing a shot change detection. Each detected shot receives a time stamp, i.e. the marking of the first shot frame in seconds. This time information can be obtained easily by knowing the frame number and the frame rate of the stream. For each shot, key frames are extracted, which will be used later on for building a “pictorial summary”, as shown in Figure 1, but also for higher-level news segmentation, as shown in Section 2. Consequently, of all time stamps collected during the shot change detection procedure, only those are kept, which correspond to starting and ending points of anchorperson shots and news shot series, which are used for building reports. The selected time stamps are then submitted to the audio processing tool, where the audio stream is segmented. Audio tracks of anchorperson shots and news shot series are then used for indexing and report forming, as explained in Section 3. For each report, the indices of its boundary shots as well as the corresponding time stamps are submitted to the multimedia database, enabling the user-interaction, as shown in Figure 1. Thereby, the shot indices are used for choosing the appropriate key frames for pictorial summaries and time stamps are required to find and retrieve the selected reports directly from the stored MPEG-compressed news program. A global technical realization scheme of the proposed news analysis, indexing and classification system is shown in Figure 5.



**Figure 5:** A global realization scheme for the proposed news analysis, indexing and classification system

In the following two sections, we will concentrate on experimental evaluation of the news segmentation, segment identification and report forming procedures, which are crucial for practical implementation of the system in Figure 5.

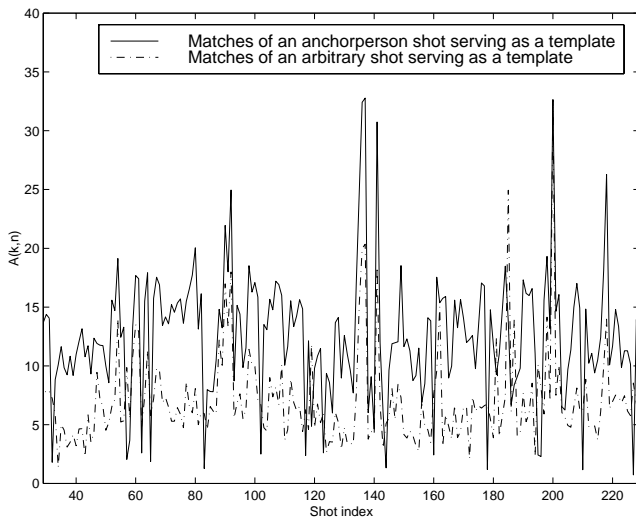
#### 4.1 The segmentation procedure

For the news program segmentation, as described in Section 2, it is required to detect all anchorperson shots and commercial breaks. For simplicity reasons, we chose the news sequences where no commercial breaks were present and concentrated only to the evaluation of the anchorperson shot detection procedure.

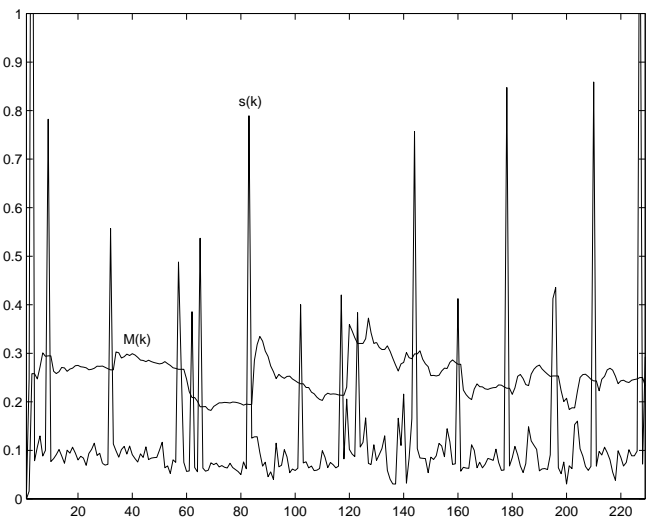
We implemented and tested our two-step detection approach using two different news sequences with following characteristics:

- *Sequence 1*: 12 minutes long, 5 anchorperson shots, one news reader, first appearance in the first sequence shot,
- *Sequence 2*: 25 minutes long, 17 anchorperson shots, two news readers, first appearance in the third sequence shot.

We represented each video shot by two subsampled key frames with sizes 165x144 for *Sequence 1* and 180x144 for *Sequence 2*. For block dimensions we chose  $M1=M2=8$ . Parameter setting for both sequences was  $N=5, P=3, \Delta k=25$  and  $w=3.0$ , whereby variations of this threshold parameter around the selected value within a comfortable interval do not influence the detection reliability. We found a 70% of all blocks in a shot image to be a good value for  $C$ . For such parameter setting we will now evaluate each of the two steps separately.



**Figure 6:** Results of the matching procedure for two different templates  $k \in [1, N]$  and shots  $[k + \Delta k, S]$



**Figure 7:** Detection diagram for Sequence 2

On both sequences we applied the template finding procedure from Section 2.2 and managed to find the proper template for each of them. Figure 6 shows the matching results of two template-candidates along the *Sequence 2*. We then measured the relative distance

$$d(m, s) = \frac{s}{m} - 1 \quad (8)$$

between the chosen minimum overall matching value  $m$  corresponding to the template, and the second smallest matching value  $s$  corresponding to the major other competitor-shot for template selection. The larger the relative distance, the more reliable is the found template. Table 2 shows these relative distances for both sequences.

Lower relative distance in the second sequence is most probably the result of the particular sequence structure showing an introduction for coming reports after the first anchorperson shot. This introduction contains very similar visual information as the shots in the later parts of that sequence, partially violating the assumption made in the introduction to this paper. However, even in such a situation, the proper template could be found, showing the robustness of the proposed dissimilarity metric.

The process of matching the found template with all the sequence shots is evaluated in Table 3 by counting missed and false detections. Two falsely detected shots in the second sequence show a dialog between the news reader in the studio and a reporter talking from a screen positioned in the corner. Obviously, the algorithm was tricked by a large amount of studio visual information and it treated these shots similarly as those showing two news readers.

	Relative distance $d(m,s)$ in percents
Sequence 1	73 %
Sequence 2	17 %

**Table 2:** Reliability evaluation of the template finding procedure

	Anchorperson shots	Detections	False alarms
Sequence 1	5	5	0
Sequence 2	17	17	2

**Table 3:** Detection results for anchorperson shots

Reliability of the detection process can be evaluated by analyzing the heights of the detection peaks in  $s(k)$  curves. One such curve, corresponding to the second sequence, is shown in Figure 7 together with the adaptive threshold  $M(k)$ .

Finally, all time stamps stored during the shot change detection procedure, and marking the beginning and the end of each anchorperson shot, are taken as the segment boundaries and used to define the same segments in the decoded audio stream. These audio segments are stored separately and used in the subsequent indexing procedure, based on wordspotting.

#### 4.2 News segment identification and report forming by wordspotting

Although being aware of more complex and robust approaches (e.g. [6, 9, 10, 13, 14]), we used a simple algorithm [18] to demonstrate the applicability of the word spotting procedure for news segment identification. The main reasons for such a choice were the availability of the software, and the simplicity of implementation and usage. The algorithm is based on template matching in frequency domain.

Since this algorithm [18] is highly speaker dependent, we had to perform the tests under idealized conditions. The idealization included the following:

- From the entire news audio sequence, only the segments spoken by one and the same anchorperson are taken into account.
- Each segment is only a couple of seconds long and contains one or two sentences on a specific topic.
- Key words for topic prespecification are extracted from the same news program and are spoken by the same anchorperson as in the first item

Such idealization is, however, not relevant, since not the actual wordspotting procedure but the usage of the wordspotting principle for topic recognition is under test. Thereby, the main issue to be tested is the efficiency of the news segment identification and report forming, as described by in Section 3 by equations (7a-b).

	Topic 1	Topic 2	Topic 3	Topic 4				Topic 5	Topic 6	
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	Segment 7	Segment 8	Segment 9	Segment10
<b>Topic 1</b>	1.122	1.369	1.347	1.292	1.282	1.333	1.356	1.372	1.292	1.440
<b>Topic 2</b>	1.401	1.282	1.349	1.442	1.324	1.371	1.506	1.465	1.457	1.559
<b>Topic 3</b>	1.783	1.801	1.647	1.878	1.717	1.844	1.855	1.818	1.855	2.015
<b>Topic 4</b>	1.338	1.330	1.339	1.254	1.222	1.337	1.397	1.351	1.319	1.480
<b>Topic 5</b>	1.359	1.453	1.389	1.388	1.355	1.410	1.457	1.476	1.300	1.539
<b>Topic 6</b>	1.739	1.854	1.828	1.816	1.817	1.838	1.811	1.914	1.757	1.311

**Table 4:** Results of key word set matching with news segments

We defined a set of 10 news segments, belonging to 6 different topics, as shown in Table 4. We also specified a set of key words (minimum two per topic) for each topic and matched each of them with every segment in order to obtain all  $W(i,j)$

values. In view of the algorithm [18], spectral energies of key words and segments obtained by short-term FFT are used for this purpose. From our experience, the longest news report does not exceed 7 segments e.g. being a combination of 4 anchorperson shots and 3 news shot series. Therefore, we recommend the setting  $G=7$  for a general case. However, due to our limited test material we will use  $G=4$  without losing generality. Let us now discuss the quality of segment identification and report forming for each of the topics.

We first find the minimum of the first row in Table 4. As expected, it lies in Segment 1. In the next step, the minimums are found in the four following columns, belonging to segments 2, 3, 4 and 5. Since neither of them is in the first row, we conclude that the Topic 1 is present only in Segment 1, which is correct. The same correct conclusions can be drawn for Topics 2, 3, 9 and 10. For Topic 4, the minimum is found in the fourth row in Table 4 in the column corresponding to Segment 5. This first result was expected since the audio sequence of this segment contains most of the words prespecified for this topic. We now find the minimums of columns 1, 2, 3, 4 and 6, 7, 8, 9, which surround the column 5. These minimums are located in the fourth row for segments 3, 4 and 8, resulting in the proper report boundary in the right side and an additional Segment 3 included in the report by mistake.

Although the used wordspotting algorithm is not robust, the matching of segments with the key word sets gives quite logical results, apart from the mistake in Segment 3 leading to the false left report boundary. All other report boundaries were found properly.

## 5. DISCUSSION

In this paper, we presented the concept of an efficient news analysis, indexing and classification system, entirely based on realistic assumptions. We also proposed practical solutions for its crucial components, being the news segmentation, segment identification and report forming procedure, which were all experimentally evaluated. We see the insufficient robustness of currently available wordspotting algorithms being the major weak point of our system. An algorithm is required which is speaker independent and having a large vocabulary. Also some other related problems need to be solved, such as the problem of different languages, dialects etc. However, we are confident that the further development in the area of wordspotting will soon result in solution which overcome the mentioned problems.

## REFERENCES

- [1] Ariki Y., Iwanari E., Montegi Y.: *Detection and Description of TV News Article*, 47th FID, pp. 198-202, 1994
- [2] Ariki Y., Saito Y.: *Extraction of TV News Articles based on Scene Cut Detection using DCT Clustering*, ICIP '96, Vol. 3, pp. 847-850, Lausanne CH, 1996
- [3] Bridle J.S.: *An Efficient Elastic-Template Method for Detecting Given Words in Running Speech*, British Acoustical Society Meeting, April 1973
- [4] Brown M.G., Foote J.T., Jones G.J.F., Sparck Jones K., Young S.J.: *Automatic Content-Based Retrieval of Broadcast News*, ACM Multimedia '95, pp. 35-43, San Francisco CA, 1995
- [5] Chen L., Faudemay P.: *Multi-Criteria Video Segmentation for TV News*, IEEE First Workshop on Multimedia Signal Processing, Princeton NJ, 1997
- [6] El Meliani R., O'Shaughnessy D.: *Accurate Keyword Spotting Using Lexical Fillers*, Proceedings of ICASSP '97, Munich, Germany 1997
- [7] ETS 300 421, Digital broadcasting systems for television, sound and data services; framing structure, channel coding and modulation for 11/12 GHz satellite services, EBU/ETSI JTC, December 1994.
- [8] Furht B., Smoliar S.W., Zhang H.: *Video and Image Processing in Multimedia Systems*, Kluwer Academic Publishers, 1995
- [9] Kanazawa H., Tachimori M., Takebayashi Y.: *A Hybrid Word Spotting Method for Spontaneous Speech Understanding Using Word-Based Pattern Matching and Phoneme-Based HMM*, Proceedings of ICASSP '95, Detroit, Michigan, 1995

- [10] Knill K.M., Young S.J.: *Speaker Dependent Keyword Spotting for Accessing Stored Speech*, Technical Report TR-193, Cambridge University, Engineering Department, 1994
- [11] Lienhart R., Kuhmuench C., Effelsberg W.: *On the Detection and Recognition of Television Commercials*, Proceedings of IEEE ICMCS '97, Ottawa, Canada, 1997
- [12] Low C.Y., Tian Q., Zhang H.: *An Automatic News Video Parsing, Indexing and Browsing System*, ACM Multimedia Conference, Boston USA, 1996
- [13] Manos A.S., Zue V.: *A Segment-Based Wordspotter Using Phonetic Filler Models*, Proceedings of ICASSP '97, Munich, Germany 1997
- [14] Suhardi, Fellbaum K.: *Wordspotting Using a Predictive Neural Model for the Telephone Speech Corpus*, Proceedings of ICASSP '97, Munich, Germany 1997
- [15] Swanberg D., Shu C.-F., Jain R.: *Knowledge guided parsing in video databases*, IS&T/SPIE Electronic Imaging, Science and Technology, San Jose, CA, 1993
- [16] The SMASH project home page: <http://www-ict.its.tudelft.nl/smash>
- [17] The STORIT project home page: <http://www-ict.its.tudelft.nl/storit>
- [18] Ward N.: *The LOTEK Speech Recognition Package*, <ftp://sanpo.t.u-tokyo.ac.jp:/pub/nigel/lotec>, 1994
- [19] Zhang H., Smoliar S.: *Developing Power Tools for Video Indexing and Retrieval*, Proceedings of IS&T/SPIE Electronic Imaging, Storage and Retrieval for Image and Video Databases II, San Jose, CA, 1994

