

Template-based Detection of Anchorperson Shots in News Programs*

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

Abstract

For easy browsing and retrieval processes in a large news database, stored news programs have to be indexed in a suitable way. One possibility of doing this is to index each news report based on covered topic. However, before the topic recognition and indexing steps can be carried out, each news program has first to be segmented into single reports. The first step in this segmentation process is the classification of all video shots of a news program in different categories, such as anchorperson (news reader) shots, news (report) shots and possible commercial breaks. In this paper we present a new reliable approach for detection of all anchorperson shots in an arbitrary news program. Increase in detection reliability, compared to the referred methods, is achieved by using a sequence-own video shot as the template for detecting all anchorperson shots of that sequence. The template itself is determined in a robust, threshold-free way.

1. Introduction

Among the types of programs to be taken into account when developing retrieval tools for modern digital video archives in different application areas, news programs appear to be important “storing objects”. The reason lies undoubtedly in their information content, which may be useful for applications in many professional areas as well as for user-private needs. Consequently, large-scale digital news archives are being created with steadily increasing volumes which are not easy to handle unless the stored information is indexed in a suitable way. Several publications can be found in literature dealing with the problem of news archive organization [1, 2, 3, 4, 5, 6, 7].

One possibility of doing this is to index each stored news report based on covered topic. At a later stage, all news reports on same or similar topics can be grouped

together resulting in a highly transparent and easy accessible news archive. However, before such an organization can be achieved, each news program has first to be segmented into single reports. The first step in this segmentation process is the classification of all video shots of a news program in different categories, such as anchorperson (news reader) shots, news (report) shots and possible commercial breaks.

In this paper we concentrate on the problem of automatically detecting *anchorperson shots* in an arbitrary news program and propose a new approach for performing this operation. Compared to already existing detection methods (e.g. [1], [2]), 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 a realistic news program structure as shown in Figure 1. Similar structure is also proposed in [2] and is more general than the one shown in [7].

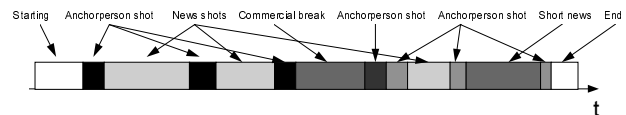


Figure 1: Assumed realistic structure of a news program. Different colors of blocks belonging to anchorperson shots indicate their different types

After making all the necessary assumptions and defining all the terms in Section 2, we present our two-step detection approach in Section 3. Section 4 defines the dissimilarity metric used to compare video shots of a news sequence. Section 5 contains the experimental results, while in Section 6 a discussion about used assumptions and their validity can be found.

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

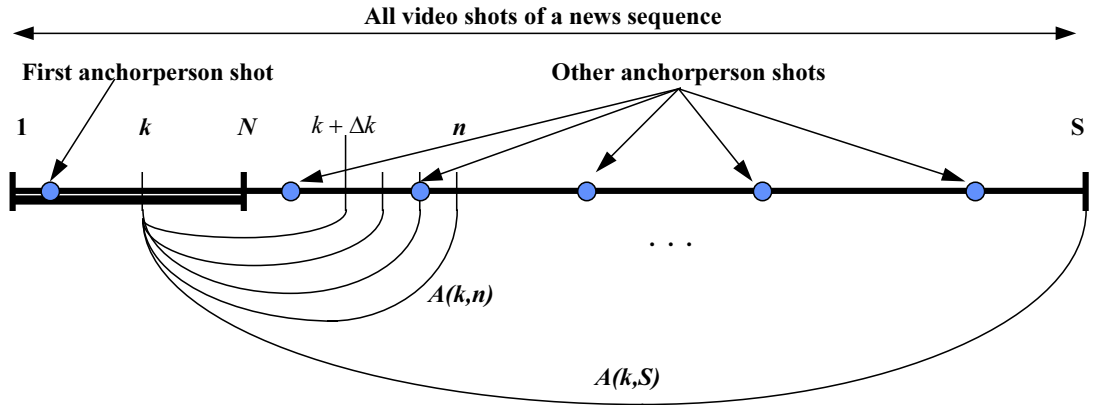


Figure 2: Obtaining of a dissimilarity values set for the shot k

2. Assumptions and definitions

We base our anchorperson detection approach on the assumption that an anchorperson shot is the only type of news sequence shots that has multiple matches of most of its visual content along the entire news program. Other (news) shots match well only in their closest neighborhood (e.g. within a single report) where they can 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 vary around 5.

In order to make the detection as robust as possible, we took into account the cases where different types of anchorperson shots appear, 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 contain certain percentage of same or similar visual features.

During the detection procedure we compare video shots based on their visual abstracts. In this way, we open

the possibility to postprocess any sequence, which has already been put in a database in the form of its visual abstract but for which no other information (e.g. temporal, audio, etc.) is available. Therefore, we assume that 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*.

3. Detection procedure

The proposed anchorperson detection approach consists of two steps, described in Sections 3.1 and 3.2, 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.

3.1. Finding a template

Different types of anchorperson shots can be found in Figure 1 together with news shots (reports) and commercial breaks. Based on assumptions from Section 2 we match each shot $k \in [1, N]$, $N < S$, with all other news shots $n \in [k + \Delta k, S]$. In this way, a set of dissimilarity values $\{A(k, k + \Delta k), \dots, A(k, S)\}$ is obtained for each shot k , as shown in Figure 2. The dissimilarity measure used here to compute values $A(k, n)$ compares two shots on basis of their abstracts (key frames) and will be explained in Section 4.

The “security” interval $[k, k + \Delta k]$ serves to avoid an possible good match of a news shot in its neighborhood. and to expose the shot k_{ap} even stronger from the rest.

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.

3.2. 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 $M(k)$, proposed in [8] and 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 $s(k) > M(k)$.

4. Dissimilarity measure

We now define the dissimilarity metric $A(k, n)$ based on [8], 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 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 in one large variable size image, called the *shot image*, which is then divided into blocks of $M \times N$ 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 $M \times N$ 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 $M \times N$ 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 3 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:

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

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:

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

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

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

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

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

where ∞ 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 all key frames of a shot k , leading to one value $d_m(b_k)$ for each block of a shot image k .

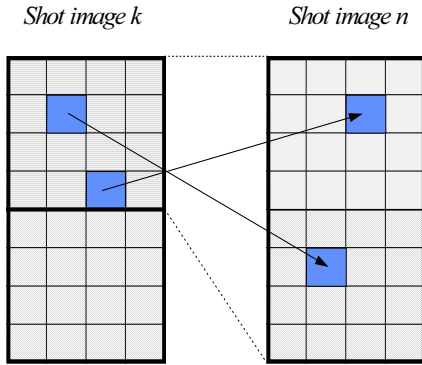


Figure 3: Comparison of a shot k with a shot n by matching $M \times N$ blocks from each key frame of shot k with all blocks in the shot image n .

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 global level,

allowing for differences among different types of anchorperson shots.

5. Experimental evaluation

We tested our two-step anchorperson shot detection approach on two different news sequences:

- *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 165×144 for *Sequence 1* and 180×144 for *Sequence 2*. For block dimensions we chose $M=N=8$. 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.

5.1. The template-finding procedure

On both sequences we applied the template finding procedure from Section 3.1 and managed to find the proper template. We then measured the relative distance

$$d(m, s) = \frac{s}{m} - 1 \quad (7)$$

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.

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

Table 1: Reliability evaluation of the template finding procedure

The larger the relative distance, the more reliable is the found template. Table 1 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.

5.2. The template-matching procedure

The process of matching the found template with all the sequence shots is evaluated in Table 2 by counting missed and false detections.

	Anchorperson shots	Detections	False alarms
<i>Sequence 1</i>	5	5	0
<i>Sequence 2</i>	17	17	2

Table 2: Detection results

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.

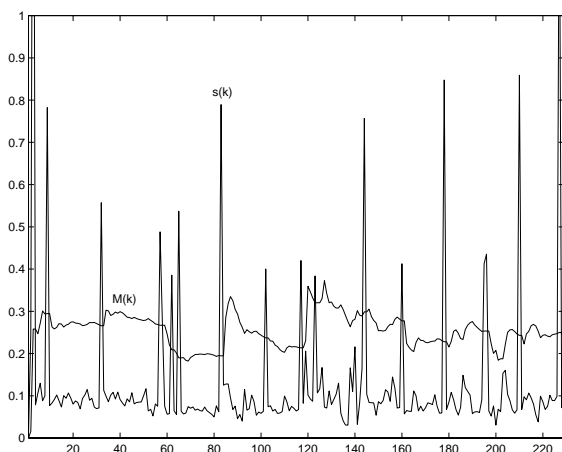


Figure 4: Detection diagram for sequence 2

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 4 together with the adaptive threshold $M(k)$.

6. Discussion

As shown by experimental results, the proposed anchorperson detection approach can perform with acceptable reliability under given assumptions. The most important assumption is that no shot of a news sequence other than anchorperson shot can find P good matches

along the entire sequence. And indeed, a small but definite probability for failure of this condition can be the major reason for lowering the algorithm's robustness in a general case, which can be observed on a lower relative distance for the second sequence in Table 1. We believe, that this problem can be solved by further improving the inter-shot dissimilarity metric in order to better distinguish different types of anchorperson shots from the rest of the sequence, while at the same time allowing for variations among these types.

7. References

- [1] Ariki Y., Saito Y.: "Extraction of TV News Articles based on Scene Cut Detection using DCT Clustering", IICIP '96, Vol. 3, pp. 847-850, Lausanne CH, 1996
- [2] Furht B., Smoliar S.W., Zhang H.: "Video and Image Processing in Multimedia Systems", Kluwer Academic Publishers, 1995
- [3] 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
- [4] Low C.Y., Tian Q., Zhang H.: "An Automatic News Video Parsing, Indexing and Browsing System", ACM Multimedia Conference, Boston MA USA, 1996
- [5] 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
- [6] Ariki Y., Iwanari E., Montegi Y.: "Detection and Description of TV News Article", 47th FID, pp. 198-202, 1994
- [7] Chen L., Faudemay P.: "Multi-Criteria Video Segmentation for TV News", IEEE First International Workshop on Multimedia Signal Processing, Princeton NJ, 1997
- [8] Hanjalic A., Legendijk R.L., Biemond J.: "Automated Segmentation of Movies into Logical Story Units", Submitted for review to IEEE Transaction of Circuits and Systems for Video Technology, Special Issue on Image and Video Processing for Emerging Interactive Multimedia Services