Inter-video Similarity for Video Parsing

Arne Jacobs, Andree Lüdtke and Otthein Herzog

Abstract In this paper we present a method for automatic detection of visual patterns in a given news video format by investigating similarities in a set of videos of that format. The approach aims at reducing the manual effort needed to create models of news broadcast formats for automatic video indexing and retrieval. Our algorithm has only very few parameters and can be run fully unsupervised. It shows good performance on a news format of the TRECVID'03 data which had already been modeled with hand-selected visual patterns and served as ground truth for evaluation.

Key words: Data Mining, Image Processing, Information Retrieval

1 Introduction and Related Work

News video broadcasts often expose a strong audiovisual and temporal structure, i.e., they are conventionalized in many ways. In most cases this structure is made pretty obvious to the viewers, for they shall be able to follow the structure of the broadcast. Paired with an audiovisual design which is characteristic to a news format this helps the viewers, on the one hand, to understand what is currently going on and, on the other hand, to recognize a certain news format. These common properties of news videos can be exploited in automatic video analysis algorithms, particularly those that focus on so-called "video parsing" [Swanberg et al(1993)Swanberg, Shu, and Jain]. In this paper we present a method to automatically detect visual patterns in a given news video format by investigating similarities between several videos of that format. It represents the logical continuation of the algorithm presented in [Jacobs(2006)] and helps in fur-

Arne Jacobs · Andree Lüdtke · Otthein Herzog
Universität Bremen, Am Fallturm 1, D-28359 Bremen, e-mail: {jarne|aluedtke|herzog}@tzi.de

ther reducing the manual effort needed to create models of news broadcast formats for automatic video indexing and retrieval.

In the next section we will describe our algorithm in detail. This is followed by experimental results which will be evaluated in Sect. 4. We conclude in Sect. 5.

2 Proposed Approach

Our goal is to find sequences in a given news broadcast format that are used to structure videos of that format and that can be used as "tokens" in a grammar for that broadcast. To achieve this goal we try to find visually near-identical subsequences that occur in most of the videos of the given format. The algorithm is based on the assumption that each news video follows a certain audiovisual design which is characteristic for its format and relatively fixed over time.

We denote a video V as a series of frames $V(t)$ at time t. We denote a subsequence of V with length l, starting at t_0 , and thus spanning the interval $[t_0, t_0 + l)$, with $V(t_0, l)$. To measure the similarity between two given subsequences we first define a similarity measure S for two single frames, where high values denote high similarity and low values denote little similarity:

$$S(V(t), V'(t')) \in [0, 1] \tag{1}$$

As we are interested in finding near-identical sequences we define a binary similarity $\dot{S}$ by applying a threshold $s_{\min}$ to our real-valued similarity measure:

$$\dot{S}(V(t), V'(t')) = S(V(t), V'(t')) \geq s_{\min} \tag{2}$$

By application of this threshold we account for slight differences between nearly identical frames caused by, e.g., different recording conditions of different videos of the same broadcast format, noise, encoding artifacts, ticker text independent of the actual news content, etc. Based on Eq. (2) we define a binary similarity measure between two video sub-sequences of the same length by the logical conjunction of the binary similarities of all temporally corresponding frames of the two sequences:

$$\dot{S}(V(t_0, l), V'(t'_0, l)) = \bigcap_{i=0}^{l-1} \dot{S}(V(t_0 + i), V'(t'_0 + i)) \tag{3}$$

To find such similar sequences we apply the following scheme: We choose a reference video V_r from the set of n videos $\{V_1, \ldots, V_n\}$ for a given news broadcast format. This is compared to every other video in the set. Given another video V' from the set we compare each frame $V_r(t)$ of the reference video with each frame $V'(t')$ of the second video. If a correspondence is found, i.e. if $\dot{S}(V_r(t), V'(t'))$ is true, we determine the longest interval $[t - a, a + b)$ for which

$$\dot{S}(V_r(t - a, a + b), V'(t' - a, a + b))$$

holds true. For each frame $V_r(t)$ we thus get a number of sub-sequences $V_r(t - a_i, a_i + b_i)$ for which a near-identical sequence $V'(t' - a_i, a_i + b_i)$ was found in V'. From these sequences we take the one with the highest similarity $S(V_r(t - a, a + b), V'(t' - a, a + b))$, which is defined as the average of the corresponding frame-wise similarities of the two sequences:

$$S(V(t_0, l), V'(t'_0, l)) = \sum_{i=0}^{l-1} \frac{S(V(t_0 + i), V'(t'_0 + i))}{l} \tag{4}$$

We also apply a constraint on the minimum length $l_{\min} \leq a + b$ of a sequence to account for the fact that for a human to recognize a characteristic sequence it has to exceed a certain length. To reduce the computational cost, we can use this minimum length constraint and only compute the similarity to every $l_{\min}$th frame of the reference video. This results from the observation that a sequence of minimum length $l_{\min}$ necessarily contains one of these frames and is thus still found by our algorithm.

We can now determine the set of frames of the reference video that belong to sequences for which we have found near-identical sequences in one or more of the $n - 1$ other videos. By using a threshold $n_{\min} \in [1, n - 1]$ on the minimum number of other videos with near-identical sequences and creating the set union of all corresponding sequences in the references video we have our result: A set of sequences from the reference video that have corresponding near-identical sequences in a specified minimum number of other videos.

2.1 Frame-wise Similarity Measure

For our algorithm to work we rely on a real-valued frame-wise similarity measure S as referenced in Eq. (1). For our purpose we use a color and texture based similarity measure as described in [Jacobs et al(2007a)Jacobs, Hermes, and Wilhelm]. It has been designed for indexing of very large (i.e., containing several millions of images) still image databases. It is sufficiently robust against noise and encoding artifacts. In contrast to [Chum et al(2007)Chum, Philbin, Isard, and Zisserman], who focus particularly on time efficient detection of nearidentical video sequences, we do not apply any sort of hashing or reverse file indexing and use a rather simple approach. However, as we only have a limited set of videos to be analyzed and as our algorithm aims at modeling news broadcast formats rather than being deployed in the actual retrieval stage, we find this approach sufficient.

2.2 Parameters of the Algorithm

The proposed algorithm has very few parameters that influence the results:

- The threshold $s_{\min}$ for frame-wise similarity (see Eq. (2))
- The minimum sequence length $l_{\min}$
- The minimum number of other videos with corresponding near-identical sequences $n_{\min}$

We believe that these parameters are intuitive and easy to select. In the following section we will demonstrate this by applying our approach to part of the TRECVID'03 set and showing some experimental results.

3 Experimental Results

The TRECVID'03 data set contains news videos from different news formats. We chose a subset of one particular news broadcast format – CNN Headline News – covering approximately three weeks with half an hour of video footage per day. We chose "CNN Headline News" because we already have additional ground truth data available for this particular format, which we will use in the evaluation section.

Our test set consists of 16 videos of "CNN Headline News" recorded on separate, mostly successive days. Each video has a length of approximately half an hour. We use the first video as reference.

As parameters of our algorithm we set $s_{\min} = 0.9$, which we found an adequate threshold for the underlying still image similarity measure to identify near-identical images. As minimum sequence length we chose $l_{\min} = 15$, which corresponds to approximately half a second. We assume that shorter sequences are not really useful for structuring a news broadcast, as they have to be recognized by the human viewers. We believe that these two parameters do not need to be tweaked and can remain fixed regardless of the news format the algorithm is applied to. We vary the third parameter $n_{\min}$ from 8, which is half of our videos, to 12 (three fourths) to test its influence on the results. In total there were 16 sequences identified by the algorithm, which are shown with manually chosen key frames in Fig. 1.

Table 1 shows a short description of the sequences and the detection results in the different runs with varying parameter $n_{\min}$.

Sequence nr. 5 – marked with * in Table 1 – contains a commercial at the end in addition to the title sequence in the first run. This was due to the fact that the same commercial appears in the reference video and also in 8 other videos of the set, always directly following the title sequence. In all other runs, sequence nr. 5 only contains the title sequence.

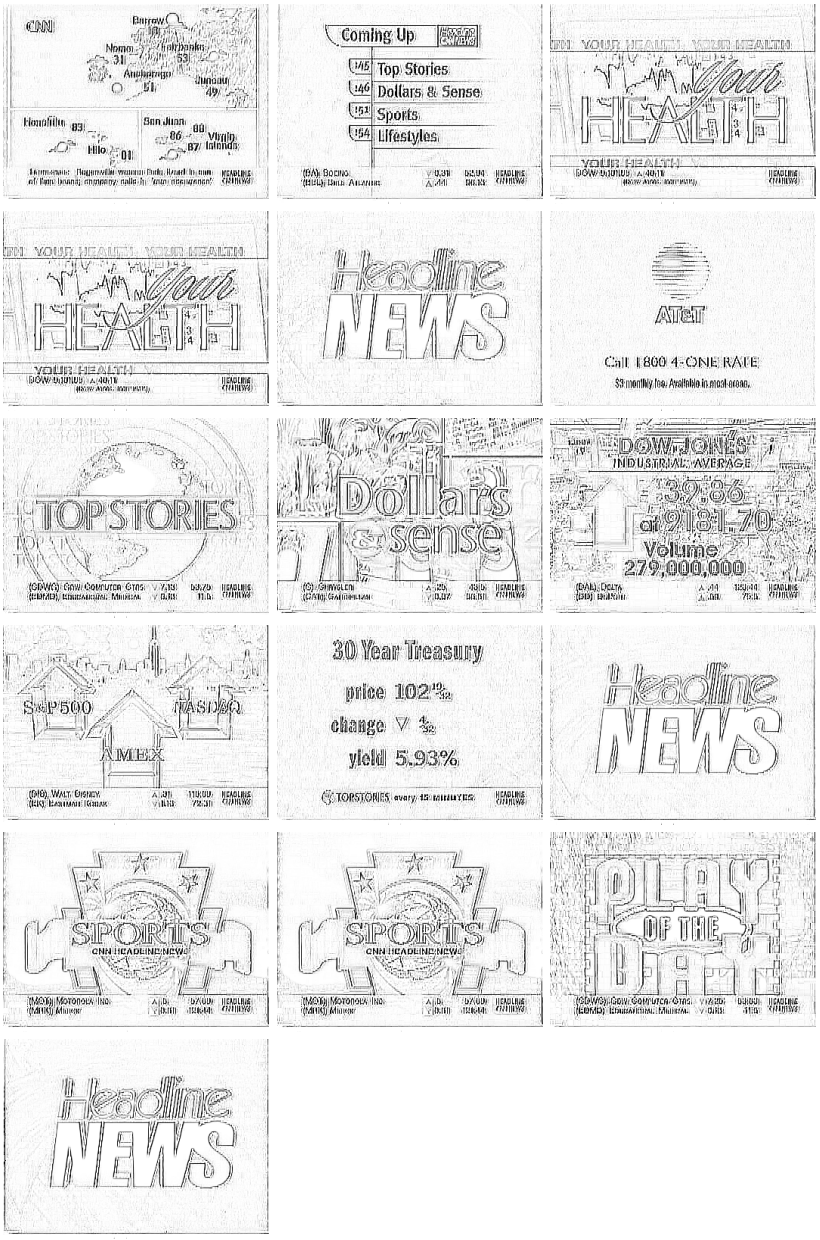

Fig. 1 Result sequences 1-16 (from top to bottom and from left to right), in order of their appearance in the reference video

Table 1 Detection results with varying $n_{\min}$, in order of their appearance in the reference video

Nr.	Description	$n_{\min}$=8	9	10	11	12
1	Weather map	X	X	X	X	
2	Coming Up screen	X	X	X	X	X
3	Your Health intro sequence	X	X	X		
4	Your Health end sequence	X	X	X	X	X
5	Headline News title sequence 1*	X	X	X	X	X
6	AT&T commercial	X	X	X	X	
7	Top Stories intro sequence	X	X	X	X	X
8	Dollars & Sense intro sequence	X	X	X	X	X
9	Financial News screen 1	X	X			
10	Financial News screen 2	X	X	X		
11	Financial News screen 3	X	X	X	X	
12	Headline News title sequence 2	X	X	X	X	X
13	Sports News intro sequence	X	X	X	X	X
14	Sports News end sequence	X	X	X	X	X
15	Play of the Day intro sequence	X	X	X	X	X
16	Headline News title sequence 3	X	X	X	X	X

4 Evaluation

For our evaluation we use the results of a project of the Delos Network of Excellence which focused on news video modelling and parsing [Jacobs et al(2007b)Jacobs, Ioannidis, Christodoulakis, Moumo
In the course of the project the "CNN Headline News" format was manually analyzed and a context-free grammar based on hand-selected visual tokens structuring the news broadcast was created by extensive examination of "CNN Headline News" example videos. The tokens finally found useful for the structural grammar were:

- "Black Frames" – a sequence of black frames
- "Channel Logo" – an animation showing the "CNN Headline News" logo
- "Coming Up" – a screen showing what is coming up next
- "Dollars and Sense Intro" – an introduction sequence to the financial news
- "Extended Forecast Map" – a weather map
- "Face" – a sequence showing a presenter in a studio setting (anchor shot)
- "Island Map" – a weather map
- "Play of the Day Intro" – an introduction sequence to the "Play of the Day" sports section
- "Pressure Map" – a weather map
- "Sports Intro" – an introduction sequence to the sports section
- "Studio" – a sequence showing an arbitrary view of the studio
- "Temperature Map" – a weather map
- "Top Stories Intro" – an introduction sequence to the "Top Stories" section
- "Your Health Screen" – an introduction/end sequence of the "Your Health" section

We can now map our automatically detected sequences to the hand-selected ones shown above. Table 1 lists the respective precision and recall values for the different runs in the second and third column. In practice, however, we find that our previous work already covers the detection of anchor shots – "Face" in the above grammar [Jacobs(2006)] – and we never expected the approach presented here to detect those sequences. By examining the grammar we can also see that all visual tokens corresponding to different weather maps always occur directly after one another. Thus it makes sense to combine them into one single token. Also, the "Black Frames" token has no real structural purpose in the grammar as it only occurs together with other structural tokens. In fact, our algorithm included the black frames into the detected sequences. The fourth column in Table 2 shows the "improved" recall values taking these observations into consideration.

Table 2 Performance of the algorithm based on manually selected ground truth sequences. R_1=Recall; P_1=Precision; R_2="Improved" recall

$n_{\min}$	R1	P1	R2
8	57%	69%	89%
9	57%	75%	89%
10	57%	80%	89%
11	57%	85%	89%
12	50%	100%	78%

5 Conclusions

An approach for automatic detection of visual structural sequences in news broadcasts was presented. It shows good performance on a standard data set which has already been modelled with hand-selected visual patterns. Structural tokens that vary greatly between several instances of a given format, e.g., anchor shots varying due to changes in presenter and/or clothing and general studio shots, are not detected by our algorithm, but this was expected.

References

[Chum et al(2007)Chum, Philbin, Isard, and Zisserman] Chum O, Philbin J, Isard M, Zisserman A (2007) Scalable near identical image and shot detection. In: Proceedings of the 6th ACM international conference on Image and video retrieval, pp 549–556

[Jacobs(2006)] Jacobs A (2006) Using self-similarity matrices for structure mining on news video. In: Proceedings of the 4th Hellenic Conference on AI SETN 2006, Heraklion, Crete, Greece, pp 87–94

[Jacobs et al(2007a)Jacobs, Hermes, and Wilhelm] Jacobs A, Hermes T, Wilhelm A (2007a) Automatic image annotation by association rules. In: Electronic Imaging & the Visual Arts EVA 2007, Berlin, Germany, pp 108–112

[Jacobs et al(2007b)Jacobs, Ioannidis, Christodoulakis, Moumoutzis, Georgoulakis, and Papachristoudis] Jacobs A, Ioannidis G, Christodoulakis S, Moumoutzis N, Georgoulakis S, Papachristoudis Y (2007b) Automatic, context-of-capture-based categorization, structure detection and segmentation of news telecasts. In: Proceedings of the First International DELOS Conference - Revised Selected Papers, Pisa, Italy, pp 278–287

[Swanberg et al(1993)Swanberg, Shu, and Jain] Swanberg D, Shu C, Jain R (1993) Knowledge guided parsing in video databases. In: Proceedings of the IS-T/SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA, vol 1908, pp 13–24