NEWS STORY SEGMENTATION IN THE FÍSCHLÁR VIDEO INDEXING SYSTEM

N. O'Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy, A. Smeaton

Centre for Digital Video Processing
Dublin City Univeristy
Ireland
Noel.OConnor@eeng.dcu.ie
http://lorca.compapp.dcu.ie/Video

ABSTRACT

This paper presents an approach to segmenting individual news stories in broadcast news programmes. The approach first performs shot boundary detection and keyframe extraction on the programme. Shots are then clustered into groups based on their colour and temporal similarity. The clustering process is controlled using the groups' statistics. After clustering, a set of criteria are applied and groups are successively eliminated in order to converge upon a set of anchorperson groups. The temporal locations of the shots in these anchorperson groups are then used to segment the programme in terms of individual news items. This work is carried out within the context of a complete video indexing, browsing and retrieval system.

1. INTRODUCTION

The Centre for Digital Video Processing at Dublin City University is pursuing an on-going research effort to develop essential technologies required for efficient management of video content. The project concentrates on fully automatic video indexing processes addressing both shot-level and scene-level video segmentation. The Centre also addresses the provision of good video content navigation and browsing support for end-users, which is considered to be an equally important aspect of video management [1].

The work of the Centre to date is demonstrated via the web-based Físchlár[1] system. Físchlár is an on-line demonstration system which allows users to (i) browse today's and tomorrow's television listings, (ii) select television programmes to be recorded, analysed and indexed, (iii) view the visual index created by the system's indexing tools and (iv) select content, based on the index, and have it streamed to them in real-time [1]. Users can select programmes from eight terrestrial public broadcast channels and television schedules can be viewed by channel, programme genre (e.g. comedy, drama, sports, etc.) or day (i.e. today or tomorrow). Most recently, a personalised listing service was introduced in order to offer programme recommendations based on user feedback on previously recorded content [2].

When a programme is recorded, it is captured in MPEG-1 format and stored on the system's video server. This MPEG-1 video bitstream is then analysed using a set of indexing tools in order to create a visual index for the content. The visual index is presented to users in one of eight different interfaces which facilitate non-linear browsing of the content [3]. In this paper, the indexing tools employed in order to create a visual index for the specific case of news programmes are described.

The visual index used in Físchlár is described in the following section. The analysis tools used to instantiate this index are outlined in section 3. Some illustrative results are presented in section 4. Finally, directions for future research are outlined in section 5.

2. A HIERARCHICAL VISUAL INDEX

The visual index used is Físchlár is designed to be be hierarchical in nature – see Figure 1. At the lowest level of the hierarchy are the results of shot-level analysis of the video content. The next level in the hierarchy contains groupings (termed clusters) of shots which are similar in terms of their visual signal-based properties. The level above this contains scene information, where a scene is defined as a collection of semantically related shots. Note that certain programmes (e.g. football matches) may not have the concept of a scene associated with them. The next level of the hierarchy contains semantic boundary information allowing the description of story lines, where a story-line is defined as a collection of different scenes which are semantically related. A further level in the hierarchy which contains references to specific objects (e.g. particular characters in a

The work described in this paper was funded by the National Software Directorate of Ireland with additional support from the Reseach Institute for Networks and Communications Engineering (RINCE).

[1]The name Físchlár is derived from two words in the Irish language: *fís* meaning dream or vision and *chlár* meaning programme.

0-7803-6725-1/01/$10.00 ©2001 IEEE

movie) and/or events (e.g. the goals in a football match) is possible.

The first two levels of the visual index can be instantiated for any type of programme using automatic visual indexing tools such as shot boundary detection and shot clustering. However, to automatically instantiate the higher levels, some knowledge of the type of content to be analysed is required. For this reason, the indexing tools employed in Físchlár to index these levels are *genre specific*. Knowledge of the genre of the programme to be indexed can provide clues to its structure which can be used to guide the analysis. In this paper we present the visual analysis tools used to instantiate the visual index in the particular case of news programmes.

A news programme typically has a very strong underlying structure and different models of this structure are possible. For the purposes of our initial experiments a simple model appropriate to the specific type of news programmes under consideration was adopted. It is assumed that each programme consists of a set of individual news stories presented by a (small number of) anchor person(s). Each story has a introduction segment consisting of the anchor person, followed by a series of reports and/or an interview segment. A story may have a closing section similar to the introduction. The different anchor persons may present alternate stories or both present the same story. The start of a new story is signalled with a new representative image or graphic somewhere in the (static) background. In the context of news programmes, only the first three levels of the Físchlár visual index are relevant whereby the scene-level of the hierarchy should be populated with individual news stories.

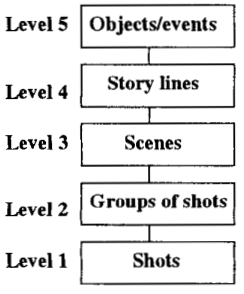


Fig. 1. The Físchlár visual index

3. INDEXING TOOLS

The generic tools used to instantiate levels 1 & 2 of the visual index are described in sections 3.1 and 3.2, whilst the genre specific tools used to instantiate level 3 are described in sections 3.3 and 3.4.

3.1. Shot boundary detection and keyframe extraction

In order to detect shot boundaries, a histrogram-based approach is employed. This approach is outlined in detail in [4, 5] and only a brief overview is presented here. A 64 bin histogram is computed for the Y, U and V components of each decoded frame. These histograms are then concatenated to form a single 192 point colour signature vector for the frame. The difference between successive vectors is calculated using the cosine similarity measure. A dynamic thresholding operation is applied to these differences in order to detect shot boundaries. Shot boundary detection approaches based on edge detection (for fades/dissolves) and MPEG-1 macroblock types (for improved computational efficiency) are also being investigated [5, 6].

Given the shot boundaries, a representative keyframe is extracted for each shot. To this end, the average colour signature (as defined above) for each shot is calculated. The keyframe is chosen as the frame in the shot whose colour signature is closest to this average calculated using the cosine similarity measure [4]. Subjectively, this approach was found to perform *marginally* better than approaches which simply select the first, last or middle frame of a shot.

3.2. Shot clustering

The shot clustering algorithm is the core component in the segmentation of news programmes. The algorithm we have implemented is based on the temporally constrained clustering approach of Rui *et al* [7]. The main difference between our approach and that of Rui *et al* is the choice of features used for each shot. We use a single feature extracted from each keyframe, rather than the multiple feature approach of Rui *et al*. We have found that this approach has worked well for our preliminary investigations but recognise that it will need to be extended in the future.

The algorithm groups shots based on the similarity of their colour composition and the temporal distance between the shots. Each shot is represented by the colour signature vector of the keyframe associated with the shot. In this way, each shot is represented as a point in a multi-dimensional feature space. Possible distance measures in this space are the Euclidean distance measure and the cosine distance measure. In this paper we have chosen to use the cosine distance measure.

The *colour similarity* measure between the shots is given by the formula:

$$ColSim(S_A, S_B) = CDM_{192}(S_A, S_B) \qquad (1)$$

where $CMD_{192}(S_A, S_B)$ is the cosine distance measure between Shot A and Shot B. The decision to place two shots

in the same group, however, depends not only on the colour similarity between them, but also on how close the temporal distance between the shots. In this way, two shots that are very similar spatially (in terms of colour composition) but are far apart in time will not be placed in the same group. To this end, the colour similarity measure is weighted by the temporal weighting function [7]:

$$TW(S_A, S_B) = max\left(0, 1 - \frac{S_A - S_B}{T_L}\right) \quad (2)$$

where S_A-S_B is the temporal distance between shot A and shot B measured in terms of the number of shots which separate them and T_L is the desired length of the temporal weighting function, again measured in number of shots. The overall shot similarity is calculated as:

$$Sim(S_A, S_B) = TW(S_A, S_B) \times ColSim(S_A, S_B) \quad (3)$$

If there are currently g groups in the clustering process, then the candidate grouping G to which to assign the current shot, S_{curr} is calculated as:

$$G = min(Sim(S_{curr}, G_i))\ \forall G_i \in G_1 \ldots G_g \quad (4)$$

Once G is found, a decision must be made as to whether or not S_{curr} can be added to that group or not. This decision is based on a *dynamic* threshold. For each group with three or more members, *the shot similarity mean*, μ, and the *shot similarity standard deviation*, σ, are calculated. The condition which must be satisfied in order for a shot to join a particular group is: $\mid \mu - sim \mid < 1.25\sigma$, where sim is the overall similarity between $shot_{curr}$ and the closest group G. When a shot is assigned to a group, the group statistics are updated and the threshold modified accordingly. If the decision is made *not* to assign S_{curr} to G, then this shot forms a new group G_{g+1}. This occurs quite frequently at the start of the clustering process until a representative set of clusters is obtained.

3.3. Anchorperson identification

The result of the shot clustering algorithm is a set of groups of shots which satisfy the conditions outlined in section 3.2. Experimental results have shown that anchorperson shots are usually assigned to a single group, in the case of a single anchorperson, or to three groups, if there are two anchorpersons.

In the majority of cases, anchorperson shots simply consist of a static background and the foreground person who reads the news. This feature ensures high similarity values between subsequent anchorperson shots throughout the news programme. To detect the anchorperson shots, the value of T_L in the temporal weighting function is set to a very long value, to avoid unnecessarilly dispersing anchorperson shots across a number of groups. This fact, in conjunction with the high threshold employed, ensures that anchorperson shots will be in the same group.

When detecting anchorperson shots there are a number of possible scenarios:

- only one anchorperson exists (the most common scenario);
- there are two anchorpersons, each taking every second story;
- there are two anchorpersons, one main anchor and one for sports stories and/or weather.

In order to decide which groups constitute anchorperson groups, it is necessary to examine all the groups and attempt to successively discount groups on the basis of criteria designed to distinguish anchorperson groups. Any groups still remaining after all the criteria have been applied are then considered anchorperson groups corresponding to one of the three scenarios outlined above.

The anchorperson identification algorithm is a *greedy* algorithm that loops through all the groups and selects those which satisfy the following conditions:

1. The range of shots of the group (i.e. the distance between the first and the last shot in terms of number of shots) is higher than a predefined value (i.e. 10), since anchorperson shots are typically widely spread throughout the news programme.
2. The group shot similarity mean is higher than a very high pre-determined value. When using the cosine distance measure, for example, the value 0.90 is used. This criterion is used because anchorperson shots should be extremely similar to each other (assuming a static background).
3. The mean anchorperson shot length should be longer than 6 seconds approx.

The thresholds used in this algorithm were empirically chosen using our sample test data set.

3.4. News story segmentation

Once anchorperson groups are identified, news story segmentation is a relatively straightforward process. One approach is to simply segment the programme on the basis of the temporal location of the start of each anchorperson shot. However, in some news programmes the anchorperson is revisited either during or at the end of a news story and very often a representative graphic is used to signal a new news story. Thus, the shots on either side of a news story boundary in an anchorperson group should be extremely similar

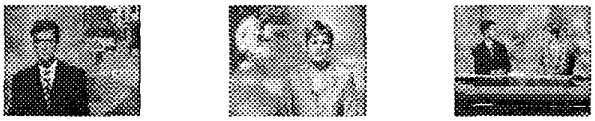
Table 1. News story segmentation results

Programme	No. stories	Precision	Recall
RTE1	14	0.93	0.81
RTE2	24	0.92	0.79
RTE3	11	0.82	0.60
RTE4	13	1.00	0.93
RTE5	10	0.90	0.90
RTE6	7	0.71	0.83
BBC1	10	0.60	1.00
Average		0.84	0.84

but not completely identical. For this reason, the similarity between successive shots in an anchorperson group is calculated. If this similarity measure is almost identical (e.g. to within 2%) then these shots are considered to be part of the same news story. News stories are then segmented as outlined above on the basis of this sub-clustering within groups.

4. RESULTS

In order to test this approach to news story segmentation a test suite was culled from the Físchlár archives. The suite consisted of both RTE (Irish national broadcaster) and BBC (English national broadcaster) news programmes. Seven programmes were chosen in all and these programmes were selected in order to be representative of the three scenarios outlined in section 3.3. Each news programme was 30 mins. in length (approx.) and featured a commercial break section. In each case, the start and end shot of each news story was marked up manually and compared to the results of the algorithm. The results obtained in terms of precision and recall are outlined in Table 1. Some illustrative results of anchor person shot detection are presented in Figure 2.



(a) (b) (c)

Fig. 2. Sample results of anchor person shot detection in a news programme

5. CONCLUSIONS

The initial results obtained with this approach to news story segmentation are very encouraging. However, the authors are aware that these results were generated on a relatively small test corpus which is not representative of the wide variety of formats of news programmes in general. Future research in this area will target a more generic approach to news story segmentation (and subsequently to scene detection in general).

To date, all analysis is carried out in the visual domain. However, other information sources should be used in order to instantiate the higher levels on the visual index. For this reason, a number of audio analysis tools are currently being developed. These include both speech vs music classification and speaker segmentation tools. In addition, it is proposed to use programme transcripts and/or teletext information (if available) for a particular programme in order to guide the news story segmentation process.

6. REFERENCES

[1] H. Lee et al, "The Físchlár digital video recording, analysis, and browsing system," in *Proc. Content-based Multimedia Information Access (RIAO'2000)*, Paris, France, 12-14 Apr. 2000.

[2] B. Smith and P. Cotter, "A personalized television listings service," *Communications of the ACM*, vol. 43, no. 8, pp. 107–111, 2000.

[3] H. Lee et al, "Implementation and analysis of several keyframe-based browsing interfaces to digital video," in *Proc. 4^{th} European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000)*, Lisbon, Portugal, 18-20 Sep. 2000.

[4] C. O'Toole et al, "Evaluation of automatic shot boundary detection on a large video test suite," in *Proc. The Challenge of Image Retrieval - 2^{nd} UK Conference on Image Retrieval (CIR'99)*, Newcastle, UK, 25-26 Feb. 1999.

[5] A. Smeaton et al, "An evaluation of alternative techniques for automatic detection of shot boundaries in digital video," in *Proc. Irish Machine Vision and Image Processing Conference (IMVIP'99)*, Dublin, Ireland, 8-9 Sep. 1999.

[6] P. Browne et al, "Evaluating and combining digital video shot boundary detection algorithms," in *Proc. Irish Machine Vision and Image Processing Conference (IMVIP'2000)*, Belfast, Northern Ireland, 31 Aug. - 2 Sep. 2000.

[7] Y. Rui, T.S. Huang, and S. Mehrotra, "Constructing a table of contents for videos," *ACM Journal of Multimedia Systems*, vol. 7, pp. 359–368, 1999.