

Split and Merge Based Story Segmentation in News Videos

Anuj Goyal, P. Punitha, Frank Hopfgartner, and Joemon M. Jose

Department of Computing Science
University of Glasgow
Glasgow, United Kingdom
{anuj,punitha,hopfgarf,jj}@dcs.gla.ac.uk

Abstract. Segmenting videos into smaller, semantically related segments which ease the access of the video data is a challenging open research. In this paper, we present a scheme for semantic story segmentation based on anchor person detection. The proposed model makes use of a split and merge mechanism to find story boundaries. The approach is based on visual features and text transcripts. The performance of the system was evaluated using TRECVID 2003 CNN and ABC videos. The results show that the system is in par with state-of-the-art classifier based systems.

1 Introduction

Processing television news has become an important and much attention seeking research work. To ease user burdens in finding segments of videos he wants, it is necessary to split videos into smaller, semantically related, segments. Identifying these smaller chunks is a real challenge. Within TRECVID, stories are defined as segments of a news broadcast with a coherent news focus which contains at least two independent declarative clauses. The main problem of text-based approaches [1,2] for story boundary detection is that textual transcripts do not always relate to the content of the actual news broadcast. Moreover, not every shot has a textual transcript. The most successful runs evaluated within TRECVID combined both low-level features and text based segmentation techniques [3,4]. There are enough evidences to understand that this area of story segmentation is still under explored and remains an open research area. Thus, in this paper, we present a scheme for story segmentation.

2 Story Segmentation

The proposed story segmentation approach comprises of two stages. The first stage proposes a method of detecting anchor person shots (APS), while the second stage suggests a sequel story boundary detection approach.

2.1 Detection of Anchor Person Shots (APS)

In order to detect anchor person shots, we use knowledge about the structure of news broadcasts, where APSs are the most similar and repeated video segments in any news bulletin. A closer analysis of the news videos in our collection reveals that the anchor person appears for the first time in the video between the time span of 25 to 55 seconds. With this backdrop, our proposed approach for detecting anchor person shots is a two step process, where the first step deals with finding an anchor person template and the second step detects all anchor person shots in the video.

Anchor person template selection. As anchor person shots retain a majority of the color distribution throughout the video, frames with abrupt change in the colour features within 25 to 55 seconds are extracted as possible candidate frames of an anchor person. Let $K_T = \{k_1^1, k_2^1, \dots, k_N^1\}$ be the frames selected as possible candidates for anchor person frames within this time span with $|K_T| = N$. Let $S = \{s_1, s_2, \dots, s_n\}$ be the set of shots in the news video and let $K = \{k_i^f, k_i^m, k_i^l \mid \forall i = 1, 2, \dots, n\}$ be the frames representing S , such that k_i^f, k_i^m, k_i^l are the first, middle and last keyframes of the shot $s \in S$. Now, in order to select an appropriate template as the anchor person template, we expect the average visual similarity of the anchor person template frame with the k_i^m frames of K to be minimal. However, although we assume the repetition of only the anchor person shots in the video and since we are using only visual features to detect the template, we skip a few shots Δk following and preceding the time span assumed to avoid false selection of the template. Thus, the algorithm for template selection is devised as in Algorithm 1.

Algorithm 1. Anchor person template selection

Require: K_T , set of candidate anchor person frames; K , set of frames of shots in S
for all $k_i^1 \in K_T$ **do**
 for all $k_j^m \in K$ **do**
 Compute visual dissimilarity between k_i^1 and k_j^m
 end for
end for
Compute average dissimilarity for each $k_i^1 \in K_T$ using the least 8 distances with the assumption that least 8 are accurate hits.
Select k_i^1 with minimum average distance as the template frame $T_k = k_i^1$.
return T_k , template anchor person frame

2.2 Story Boundary Detection

Anchor person shot (APS) detection. Once the template anchor person frame is selected, the APS detection becomes a simple process of template matching. However to avoid possible missed detection due to the incapability of a keyframe being a real representative of a shot, we use three frames per shot for matching. Thus, the algorithm for APS detection is as given in Algorithm 2.

Algorithm 2. Anchor person shot detection**Require:** T_k , template anchor person frame; K , set of frames of shots in S $S_A = \{\}$ **for all** $k_i \in K$ **do** Compute visual dissimilarity d_i between $k_i \in s_i$ and T_k using $d_i = \min(\text{dis}(k_i^f, T_k), \text{dis}(k_i^m, T_k), \text{dis}(k_i^l, T_k))$, where dis is the dissimilarity computation function**end for**Pick all shots s_i as APSs if $d_i < T_h$, where T_h is a threshold and update S_A such that $S_A = S_A \cup \{s + i\}$.**return** $S_A = \{s_i \in S \mid S_A \subset S\}$

Despite the assumption that any story starts with the anchor person, it is not always true that a story ends with appearance of the next APS. For instance, within an APS there is a possible presence of a story boundary, as the anchor person continues with the previous story and changes to the new story only towards the mid of the APS. It could also happen that an anchor person introduces stories without any supporting video clips. This gives rise to possible, intra shot story boundaries, i.e. possibly more than one story covered within an APS or inter shot story boundaries, where an APS may or may not start a new story. Hence it is required to split or merge APS accordingly.

We start with the basic assumption that every APS begins a new story. Now let $s_i^k = \{f_1, f_2, \dots, f_d^i\}$ be a set of frames extracted for every second from shot $s_i \in S_A$. To find possible story boundary within s_i , the frames f_j and f_{j+1} are split into four regions, with R_1 and R_2 being first and second quadrants and the eigen difference (which is well accepted in the field of Face Recognition) E_1 and E_2 of R_1 and R_2 regions are computed respectively. If $E_1 < T_d$ and $E_2 > T_S$ or $E_2 < T_d$ and $E_1 > T_S$ (say, C_1) then f_{j+1} is marked to begin a new story where T_d and T_S are thresholds set to declare two images as similar or dissimilar. In any other case the frames are assumed to be the members of the same story and the process is continued with the next two frames f_{j+1} and f_{j+2} in sequence. If there is no intra story boundaries detected in s_i and s_i happens to be the first APS then the beginning of this shot starts a new story. If the shot s_i is not the first APS, then the last frame f_d^{i-1} of the preceding APS s_{i-1} provided s_{i-1} has intra story segments, and the first frame f_1^i of the shot s_i is used to compute the eigen difference between the R_1 and R_2 regions of frames f_d^{i-1} and f_1^i . If the region differences do not satisfy the condition C_1 then the last story segments of s_{i-1} and s_i are merged together as one single story. On the other hand if s_{i-1} has no intra story boundaries, then s_i is marked as the beginning of a new story. This process is carried out for all shots chosen as APSs.

2.3 Classifier Based Story Segmentation

We tested the suitability of SVM, ANN, J48 decision tree and Naïve Bayes classifiers to detect the APSs with empirically selected parameters. To classify a

shot as an APS the neighbouring shots on both hands were used with a fixed window size as the region of support. Correlation-based feature subset selector [5] and best first search method were used to select and weight a few out of many features extracted. The final selected features used for the training in order of preference are: **Distance from $T_k(D_{S_i \rightarrow T_k})$** : The distance in MPEG7 colour structure feature between T_k and s_i . **Semantic Text Similarity (S_{LR})**: The similarity [6] between the transcript of left region of support and right region of support. **Shot length difference (D_{LR})**: The absolute difference in the number of frames in left region of support to right region of support. **Average visual dissimilarity (AVG_{LR})**: The average of colour structure distance between the left region of support and right region of support. **Minimum visual dissimilarity (MIN_{LR})**: The minimum colour structure distance between shots from left region of the support to the shots from right region of the support. Thus we used the tuple $(D_{s_i \rightarrow T_k}, S_{LR}, D_{LR}, AVG_{LR}, MIN_{LR})$ to train different classifiers. These trained classifiers were then used to detect APS. To find story boundaries out of these classified APSs, we use the same algorithm described in Section 2.2.

3 Evaluation

We used the TRECVID 2003 test collection to evaluate our segmentation approach. The corpus consists of 52 hours of news videos. It contains roughly 3000 story boundaries which have been manually annotated. The transcripts provided within TRECVID and MPEG7 colour structure features from frames were used for our evaluation. Threshold values are kept fixed for both CNN and ABC videos.

The threshold based method is as good as the classifier based systems with respect to the recall as evident from Figure 1. However, there is drastic variation in the precision, reporting that the false detection of story boundaries is

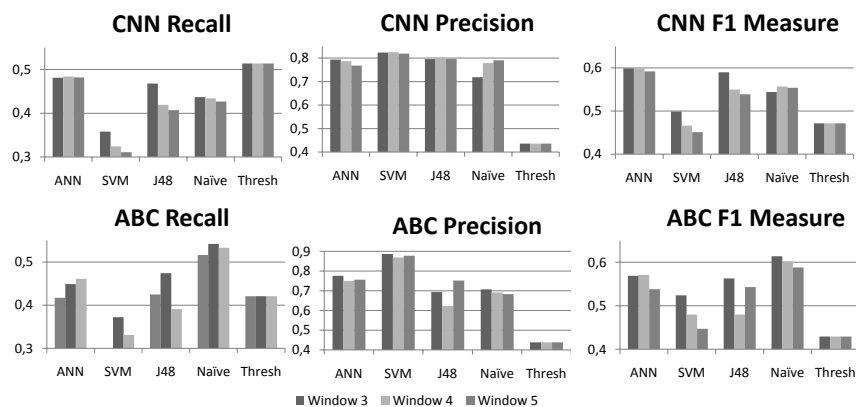


Fig. 1. Evaluation Results of Five Methods

high. Since the threshold based system completely relies on visual features of the APS, without any semantics derived from the neighbouring shots, there are possibilities of detecting false APSs. Since the story segmentation approach processes only the APSs for semantic story boundaries, non-APSs tend to form a new story by itself. This results with a drop in precision.

4 Discussion and Conclusion

The story segmentation approach proposed in this paper makes use of some heuristics made in compliance to the video structures. Though all news channels are diverse and follow different production criteria, the commonality amongst all news videos assumed throughout this paper is the presence of an anchor person. Hence the heuristics and the method revolve around finding accurate anchor person shots followed by splitting and merging. The performance of the system is in par with the state-of-art approaches which have been evaluated within TRECVID. A possible way to improve the results is to identify channel dependent threshold values. This is currently being investigated.

Acknowledgments

This research was supported by the European Commission under the contracts FP6-027026-K-SPACE and FP6-027122-SALERO.

References

1. Passonneau, R.J., Litman, D.J.: Discourse segmentation by human and automated means. *Comput. Linguist.* 23(1), 103–139 (1997)
2. Galley, M., McKeown, K., Fosler-Lussier, E., Jing, H.: Discourse segmentation of multi-party conversation. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 562–569. Association for Computational Linguistics (2003)
3. Arlandis, J., Over, P., Kraaij, W.: Boundary error analysis and categorization in the TRECVID news story segmentation task. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) *CIVR 2005*. LNCS, vol. 3568, pp. 103–112. Springer, Heidelberg (2005)
4. Chua, T.S., Chang, S.F., Chaisorn, L., Hsu, W.: Story boundary detection in large broadcast news video archives: techniques, experience and trends. In: *MM 2004*, pp. 656–659. ACM, New York (2004)
5. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: *ICML 2000*, pp. 359–366. Morgan Kaufmann Publishers Inc., San Francisco (2000)
6. Kolb, P.: DISCO: A Multilingual Database of Distributionally Similar Words. In: *KONVENS 2008 – Ergänzungsband: Textressourcen und lexikalisches Wissen* (2008)