CUNI at TRECVID 2015 Video Hyperlinking Task

Petra Galuščáková[1], Michal Batko[3] Martin Kruliš[2], Jakub Lokoč[2], David Novák[3], and Pavel Pecina[1]

[1] Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{galuscakova,pecina}@ufal.mff.cuni.cz
[2] Charles University in Prague, Faculty of Mathematics and Physics
SIRET Research Group
{krulis,lokoc}@ksi.mff.cuni.cz
[3] Masaryk University, Faculty of Informatics, Brno
{batko,david.novak}@fi.muni.cz

Abstract. In this paper, we present our approach used in the TRECVID 2015 Video Hyperlinking Task [13]. Our approach combines text-based similarity calculated on subtitles, visual similarity between keyframes calculated using Feature Signatures, and preference whether the query and retrieved answer come from the same TV series. All experiments were tuned and tested on about 2500 hours of BBC TV programmes.
Our *Baseline* run exploits fixed-length segmentation, text-based retrieval of subtitles, and query expansion which utilizes metadata, context, information about music and artist contained in the query segment and visual concepts. The *Series* run combines the *Baseline* run with weighting based on information whether the query and data segment come from the same TV series. The *FS* run combines the *Baseline* run with the similarity between query and data keyframes calculated using Feature Signatures. The *FSSeriesRerank* run is based on the *FS* run on which we applied reranking which, again, uses information about the TV series. The *Series* run significantly outperforms the *FSSeriesRerank* run. Both these runs are significantly inferior to our *Baseline* run in terms of all our reported measures. The *FS* run outperforms the *Baseline* run in terms of all measures but it is significantly better than the *Baseline* run only in terms of the MAP score. Our test results confirm that employment of visual similarity can improve video retrieval based on information contained in subtitles but information about TV series which was most helpful in our training experiments did not lead to further improvements.

1 Introduction

The Video Hyperlinking Task deals with retrieval of video segments from a video collection. The retrieved segments should be topically related to given query video segments. But instead of just being similar to the query segment, retrieved segments should give more information about the query segment. The

main objective of the task is to explore methods which enable users to easily browse the video collection using hyperlinks provided for the segments of their interests.

The data collection provided for this task consists of TV programmes created by BBC and broadcasted on BBC between May 12, 2008 and July 31, 2008. It includes high-quality videos containing a large variety of topics, locations, and persons. The videos were provided with subtitles, metadata and information on keyframes. The task was evaluated using crowdsourcing on a total of 100 queries (selected from a set of 135 queries) defined by media professionals. The results are reported using the following measures: MAP, Precision at 10 (P10), MAP-bin, MAP-tol [1], and MAISP [13].

In our experiments, we exploited the available subtitles, metadata (title and description), and keyframe information. Our experiments were tuned on 30 queries used for evaluation in the MediaEval 2014 Search and Hyperlinking Task [5].

2 Baseline System

The core of our system is similar to our set-up used in the MediaEval 2014 Search and Hyperlinking Task experiments [6]. All recordings were first segmented into 60-second passages with new passages being created every 10 seconds. We used transcripts of the passages created from available subtitles and concatenated them with corresponding metadata of the video file. We used a title and a description of the file, and the information about the broadcast channel, which we mined from the filename. These concatenated passages were indexed using the Terrier information retrieval system [12]. For the retrieval we used the Hiemstra Language Model [9] with its parameter set to 0.35. We also used Porter stemmer and Terrier's stopwords list.

Similar to data segments, queries were created from the subtitles by using all words lying inside the query segment. The queries were then concatenated with metadata of the source video file. We also used a context of the query segment – the boundaries of each query segment were enlarged by 20 seconds. The length of the context was tuned on the MediaEval 2014 Search and Hyperlinking Task training data. Compared to last-year's system, the query segments were further expanded by audio and visual information.

The BBC collection also contains concerts (e.g. *Radio 1's Big Weekend*, *Glastonbury The Best Bits*) and music programmes (e.g. *Mad about Music*, *Later... with Jools Holland*) but our system, based mainly on text retrieval, is not intended to handle information in these programmes very well. Therefore, the query segments were also expanded by audio information contained in each segment [7]. Each segment was divided into 10-second sub-segments with a new sub-segment being created each second. If the segment contains any music, the created sub-segments should be long enough to enable its recognition, while other noise and speech can be possibly cut-off. Sub-segments were then submit-

ted to the Doreso service[4] for music identification which uses a fingerprinting technique [4]. If the sub-segment contains any music, the song title and the artist are retrieved and this information is concatenated with the query segment. Music was detected in 10 out of 30 training queries but in only 7 out of 135 test queries (e.g. the query 96 contains songs *Cassava* by *Triclops!* and *Safari* by *John Barry*, and the query 69 contains the song *Something To Talk About* by *Badly Drawn Boy*).

Each query was further expanded by visual concepts contained in the source video segment. Concepts were recognized in each keyframe of query segments using the system used in the ImageCLEF 2014 Task [3]. Each concept has an associated confidence score that was used to weight individual terms in the concatenated query. Each concept was only used if it occurred in less than 7 segment keyframes, so that frequently used terms with low information value were filtered out. The number of keyframes was tuned on the training data on which this method had previously proved to be helpful.

3 System Tuning

3.1 *Series* Run

The *Series* run achieved the highest improvement on the training data. It combines the *Baseline* run with information on whether the videos came from the same TV series. This information was then used to precalculate a "series weight" for each video which was set to 0.13 if the query video and data video were from the same TV series; otherwise the weight was set to -0.15. The series weights were calculated on the training data. For each training query, we calculated the precision of retrieved results from the same TV series and from the different TV series. The average precision for the same TV series is 0.589 (the relative improvement is 13%), and the precision for the different TV series is 0.438 (the relative deterioration is 15%). Weights were calculated for each query and they were linearly combined with the top 1000 retrieved results directly in the Terrier framework. The combination weight was tuned and set to 35.

Similar to these weights, we also experimented with "time differential weights" determined by the time difference between the data and query video dates. The precision for the videos broadcast up to one week from the query video varies from 59.33 (74% of the relative improvement) in the case that the data video was broadcast up to one day from the query video to 34.574 in the case when the data video was broadcast up to one week from the query video. The average value of precision across all date differences is 34.093. The videos broadcast before the query segment have higher average precision values (35.777) than the videos broadcast after the query segment (31.598). We precalculated time differential weights based on the relative improvement of the time calculated on the training data and combined them with the *Baseline* system in the same way as that we used for TV series weights. However, the approach which favors videos from the same TV series achieved a greater improvement on the training data.

[4] `http://developer.doreso.com`

3.2 *FS* Run

The *FS* run is similar to our highest ranked run submitted to the MediaEval 2014 Benchmark. In this run, the *Baseline* was expanded by the visual similarity between the query segment and data segments. Visual similarity between each pair of keyframes contained in the query segment and data segment was calculated using Feature Signatures and Signature Quadratic Form Distance. The similarity between the segments was then calculated as the sum of similarity between the two most similar keyframes. This similarity measure was tuned on the training data. We also examined similarity calculated as an average similarity over all keyframes in the segment, maximum similarity between single keyframes, similarity between three most similar keyframes and similarity in the case when the neighbouring keyframes are enough similar/dissimilar.

Precalculated similarity between segments was then linearly combined with the score of the text-based retrieval. This combination was computed for the top 1000 retrieved results. The combination weight was tuned and set to 90. We experimented with the number of retrieved results which should be combined with visual similarity but discovered that this decreased the scores. Moreover, we experimented with mean and max-min normalization used in this combination but it did not prove to be helpful either.

3.3 Feature Signatures

In order to represent keyframes, we employ Feature Signatures that approximate distribution of color and texture in the image. Unlike modern CNN descriptors excellent in recognizing specific objects, this traditional descriptor can be used to identify keyframes with a similar background.

Formally, given a feature space $\mathbb{F}$, the *Feature Signature* S^o of a multimedia object o is defined as a set of tuples $\{\langle r_i^o, w_i^o \rangle\}_{i=1}^n$ from $\mathbb{F} \times \mathbb{R}^+$, consisting of representatives $r_i^o \in \mathbb{F}$ and weights $w_i^o \in \mathbb{R}^+$. The distance between Features Signatures is calculated using the Signature Quadratic Form Distance.

Definition 1 (Signature Quadratic Form Distance) *Given two feature signatures* $S^o = \{\langle r_i^o, w_i^o \rangle\}_{i=1}^n$ *and* $S^p = \{\langle r_i^p, w_i^p \rangle\}_{i=1}^m$ *and a similarity function* $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ *over a feature space* $\mathbb{F}$*, the Signature Quadratic Form Distance* SQFD_{f_s} *between* S^o *and* S^p *is defined as:*

$$\mathrm{SQFD}_{f_s}(S^o, S^p) = \sqrt{(w_o \mid -w_p) \cdot A_{f_s} \cdot (w_o \mid -w_p)^T},$$

where $A_{f_s} \in \mathbb{R}^{(n+m)\times(n+m)}$ *is the similarity matrix arising from applying the similarity function* f_s *to the corresponding feature representatives, i.e.,* $a_{ij} = f_s(r_i, r_j)$*. Furthermore,* $w_o = (w_1^o, \ldots, w_n^o)$ *and* $w_p = (w_1^p, \ldots, w_m^p)$ *form weight vectors, and* $(w_o \mid -w_p) = (w_1^o, \ldots, w_n^o, -w_1^p, \ldots, -w_m^p)$ *denotes the concatenation of weight vectors* w_o *and* $-w_p$*.*

Specifically, we have utilized position-color-texture Feature Signatures [10, 14] that approximate a distribution of color and texture in each keyframe. This

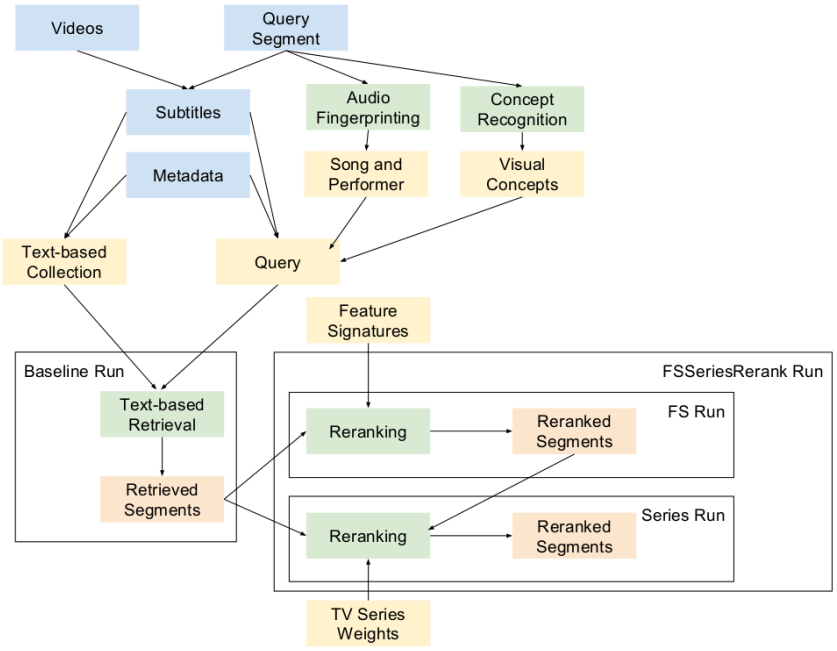

Fig. 1. Retrieval processing system - strategy diagram.

descriptor can be utilized in image retrieval tasks, where color and texture is meaningful for retrieval. As the Signature Quadratic Form Distance is a ptolemaic metric, metric/ptolemaic indexing techniques can be utilized for efficient retrieval [2, 8]. Furthermore, existing GPU implementations enable both efficient extraction of Feature Signatures and also efficient evaluation of the Signature Quadratic Form Distance [10, 11].

Employing Feature Signatures enables visual similarity to work exceptionally well in detecting similar settings and backgrounds and thus could be helpful for finding related content. This is particularly important in working with TV collections, in which a similar background occurs throughout the series. However Feature Signatures can fail to detect some details in keyframes, e.g. it is not possible recognise a particular person.

3.4 *FSSeriesRerank* Run

Finally, the *FSSeriesRerank* run is a combination of the *FS* and the *Series* runs. The top 1000 results returned by the *FS* system were linearly combined with the same weights as those used in the *Series* run and reranked accordingly. Again we experimented with several reranking scenarios but they did not improve our results. In the *FS* and *FSSeriesRerank* runs no segments were retrieved for the query 118. For this query, the answers from the *Baseline* run were used. All our experiments were tuned for the highest MAP-tol measure.

Table 1. Comparison of results submitted to the Video Hyperlinking Task. Best results for each measure are highlighted.

Run num.	Run name	MAP	P10	MAP-bin	MAP-tol	MAISP
1	Series	0.1312	0.2600	0.1443	0.1131	0.1204
2	FSSeriesRerank	0.0987	0.1980	0.1094	0.0838	0.0984
3	FS	**0.1441**	**0.2750**	**0.1560**	**0.1234**	**0.1311**
4	Baseline	0.1405	0.2740	0.1536	0.1214	0.1296

Table 2. Comparison of the postfiltered results of the Video Hyperlinking Task with query video segments filtered out. Best results for each measure are highlighted.

Run num.	Run name	MAP	P10	MAP-bin	MAP-tol	MAISP
1	Series	0.1971	0.4313	0.2005	0.1718	0.1609
2	FSSeriesRerank	0.1553	0.3414	0.1573	0.1340	0.1322
3	FS	**0.2131**	**0.4545**	**0.2138**	**0.1846**	**0.1739**
4	Baseline	0.2095	0.4495	0.2118	0.1826	0.1720

4 Results

The strategy diagram of our retrieval processing system is displayed in Figure 1. A performance comparison is reported in Table 1. The *FS* run outperformed the *Baseline* run in terms of all reported measures but the improvement is statistically significant[5] only in terms of the MAP measure. Weighting using information about the same TV series did not outperform the *Baseline* run in any case; the *Baseline* run is significantly better in terms of all scores despite its improvement on the training data.

We did not filter out segments retrieved from the query video file in the submitted results. These segments were judged as incorrect during the evaluation process, which decreased our scores. After submitting the official runs, we filtered out these segments and recalculated the evaluation scores. The recalculated results are displayed in Table 2. The postfiltered results are in all cases significantly better than corresponding original results. The best results are again achieved in the case of the *FS* run in terms of all reported measures.

Due to the nature of Feature Signatures, the visual similarity is especially helpful in the retrieval of similar backgrounds and settings. This can be used to advantage in our collection of TV programmes, in which similar backgrounds often occur in various episodes of the same TV series. Furthermore, we would like to compare and combine the Feature Signatures with other methods intended to recognize image details and faces.

Acknowledgments

This research is supported by the Czech Science Foundation, grant number P103/12/G084, Charles University Grant Agency GA UK, grant number 920913, and the SVV project number 260 224.

[5] Wilcoxon signed rank test [15] at the 0.05 level was used.

References

1. Aly, R., Eskevich, M., Ordelman, R., Jones, G.J.F.: Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. CoRR abs/1312.1913 (2013)
2. Beecks, C., Lokoč, J., Seidl, T., Skopal, T.: Indexing the Signature Quadratic Form Distance for Efficient Content-based Multimedia Retrieval. In: Proc. of ICMR. pp. 24:1–24:8. ACM, Trento, Italy (2011)
3. Budíková, P., Botorek, J., Batko, M., Zezula, P.: DISA at ImageCLEF 2014: The Search-based Solution for Scalable Image Annotation. In: Proc. of CLEF 2014 Evaluation Labs and Workshop. pp. 1–12. Sheffield, UK (2014)
4. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A Review of Algorithms for Audio Fingerprinting. In: Proc. of 2002 IEEE Workshop on Multimedia Signal Processing. pp. 169–173. IEEE (2002)
5. Eskevich, M., Aly, R., Racca, D.N., Ordelman, R., Chen, S., Jones, G.J.F.: The Search and Hyperlinking Task at MediaEval 2014. In: Proc. of MediaEval. Barcelona, Spain (2014)
6. Galuščáková, P., Kruliš, M., Lokoč, J., Pecina, P.: CUNI at MediaEval 2014 Search and Hyperlinking Task: Visual and Prosodic Features in Hyperlinking. In: Proc. of MediaEval. Barcelona, Spain (2014)
7. Galuščáková, P., Pecina, P.: Audio Information for Hyperlinking of TV Content. In: Proc. of SLAM. pp. 27–30. Brisbane, Australia (2015)
8. Hetland, M.L., Skopal, T., Lokoč, J., Beecks, C.: Ptolemaic Access Methods: Challenging the Reign of the Metric Space Model. Information Systems 38(7), 989–1006 (2013)
9. Hiemstra, D.: Using Language Models for Information Retrieval. Ph.D. thesis, University of Twente, Enschede, Netherlands (2001)
10. Kruliš, M., Lokoč, J., Skopal, T.: Efficient Extraction of Feature Signatures Using Multi-GPU Architecture. In: MMM (2). LNCS, vol. 7733, pp. 446–456. Springer (2013)
11. Kruliš, M., Skopal, T., Lokoč, J., Beecks, C.: Combining CPU and GPU Architectures for Fast Similarity Search. Distributed and Parallel Databases 30(3-4), 179–207 (2012)
12. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proc. of ACM SIGIR'06 Workshop on Open Source Information Retrieval. pp. 18–25. Seattle, Washington, USA (2006)
13. Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A.F., Quéenot, G., Ordelman, R.: TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In: Proc. of TRECVID 2015. NIST, USA (2015)
14. Rubner, Y., Tomasi, C.: Perceptual Metrics for Image Database Navigation. Kluwer Academic Publishers, Norwell, MA, USA (2001)
15. Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6), 80–83 (1945)