Multimodal Synchronization of Image Galleries

Maia Zaharieva[1,2] Michael Riegler[3] Manfred Del Fabro[4]

[1]Interactive Media Systems Group, Vienna University of Technology, Austria
[2]Multimedia Information Systems Group, University of Vienna, Austria
[3]Media Performance Group, Simula Research Laboratory AS, Norway
[4]Distributed Multimedia Systems Group, Klagenfurt University, Austria
maia.zaharieva@tuwien.ac.at, michael@simula.no, manfred.delfabro@aau.at

ABSTRACT

This paper describes our contribution to the MediaEval 2014 task on the Synchronization of multi-user Event Media (SEM). We propose two multimodal approaches that employ both visual and time information for the synchronization of different images galleries and for the detections of sub-events. The methods prove robustness in the determination of time offsets with accuracy of up to 87%.

1. INTRODUCTION

A multifaceted view of a social event can emerge when different people capture different perspectives of the same event and a compilation of all images is created. While it is typically easy to get an overview of a single image gallery, it is much more difficult to synchronize the content of two or more collections. In general, there is no guarantee that timestamps, location information or textual descriptions associated with images are correct.

In our contributions to the SEM task [1] we first focus on global visual features to identify highly similar images across different galleries of the dataset. Following, we apply visual- and time-based methods for the synchronization of galleries and for the detection of sub-events. Our first approach relies on the pairwise comparison of images in order to link different galleries. Agglomerative Hierarchical Clustering (AHC) is applied in order to group image pairs to sub-events. The synchronization offsets are calculated by iterating through the image pairs in a transitive way. In our second approach all images are clustered using the XMeans algorithm in order to identify sub-events. The synchronization offsets are estimated by calculating average time differences within the clusters.

2. APPROACHES

2.1 AHC-based Approach

We employ AHC for both time offset calculation and sub-event detection. We first cluster all images of the dataset using the MPEG7 Color Structure Descriptor (MPEG7-CS). At the very lowest hierarchy level clusters of visually highly similar images are generated. We sort these pairs of images in ascending order according to their dissimilarity level. We consider such pairs of images identical if: 1) the images originate from different galleries and 2) the dissimilarity distance does not exceed a predefined threshold. Images, representing different galleries, are considered as entry points for the synchronization of the corresponding gallery. We process the sorted pair list until we are able to build a transitive list of entry points for all galleries presented in the full dataset or we reach the end of the list. Eventually, all galleries are time aligned according to the provided reference collection using the corresponding entry points.

A higher hierarchy level of AHC already provides a reliable base for visual-based detection of sub-events. In order to avoid the building of broad clusters, we employ a strict cutoff threshold in combination with the Ward method [4] to automatically define the number of clusters. We reduce the resulting over-segmentation of underlying events by employing an adaptive, time-based approach for cluster merging. Two clusters are merged if they share a common gallery and the minimum time distance between the corresponding images is lower than a predefined threshold.

2.2 XMeans-based Approach

For this approach we employ a modified version of the algorithm presented in [3]. We select the best global feature for the given dataset by considering the information gain. The calculation is done for 13 different features (Color and Edge Directivity Desciptor (CEDD), Fuzzy Color and Texture Histogram (FCTH), Joint Composite Descriptor (JCD), Pyramid Histogram of Ortented Gradients (PHOG), Edge Histogram (EH), Color Layout (CL), Gabor, Tamura, Luminance Layout (LL), Opponent Histogram (OH), JPEG Coefficent Histrogram (JPEGCoeff), Scaleable Color (SC) and Auto Color Correlogram (ACC) [2]). JCD had the highest information gain for the SEM dataset and, therefore, it was employed for this approach.

In order to synchronize the dataset, we first cluster all images using the XMeans algorithm. Following, we consider the average deviation of the reference image timestamps to all other images of a collection that share a common cluster as offset for this image collection. If there are less than two reference images in a cluster, we use the available corrected timestamp of non-reference images which already have an offset from another cluster. For sub-event detection, we employ XMeans clustering using JCD or the corrected capture times as features.

3. EXPERIMENTS AND RESULTS

The SEM development dataset contains 304 Flickr images from the *London Olympic Games 2012*. The images are arranged in 10 galleries and represent 59 sub-events in total.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Sub-event detection results on the development dataset in terms of number of detected clusters (C), F1-score (F1), and Normalized Mutual Information (NMI).

	C	F1	NMI
Time-based clustering	98	0.6363	0.8696
AHC + MPEG7-CS	91	0.5543	0.8179
AHC + MPEG7-CS + Time	45	0.6303	0.7927
Xmeans + JCD	89	0.5123	0.7812
Xmeans + Time	100	0.5731	0.8231

Table 2: Official runs configurations.

	Time Offset	Sub-events detection
run 1	AHC + MPEG7-SC	AHC + MPEG-7 SC
run 2	AHC + MPEG7-SC	Time-based
run 3	XMeans + JCD	XMeans + JCD
run 4	XMeans + JCD	XMeans + Time
run 5	AHC + MPEG7-SC	XMeans + Time

Experiments on the development dataset show significant differences in the precision of detected time offsets between the two approaches. While, the AHC-based approach in combination with MPEG7-CS achieves 18.5 seconds deviation in average over the 10 galleries, the XMeans-based and the JCD feature obtain only 2216.4 seconds in average.

Additionally, we compare the performances of purely time-based clustering (after considering the time offsets), visual-based clustering, and the combination thereof using the AHC approach. We measure the performance by means of harmonic mean (F1-score) of recall and precision and Normalized Mutual Information (NMI) measuring the goodness of clustering of retrieved events. The results achieved show that both the time-based and the visual-based clustering result in over-segmentation of the underlying events (90+ detected sub-events vs. 59 ground truth events) and high NMI scores. The combination of visual and time information outperforms the visual-based approach and significantly reduces the number of detected sub-event clusters (see Table 1). Noteworthy is the observation that with both approaches, the time-based detection of sub-events outperforms the corresponding visual-based approach in terms of F1 (at higher over-segmentation costs).

We submitted five runs for the final evaluation (see Table 2 for the configurations). Tables 3 and 4 summarize the corresponding results for the synchronization and for the sub-event detection task. Results on the synchronization task are reported in terms of precision (percentage of synchronized galleries with a misalignment lower than 30 minutes), and accuracy (closeness of detected offset to real offset, normalized with respect to the maximum accepted time lapse of 30 minutes). The results achieved confirm our experiments on the development dataset: the AHC-based approach in combination with the MPEG-7-CS clearly outperform our XMeans-based approach. Although both datasets contain approximately the same number of galleries (35 Vancouver, 37 London) they perform differently. The Vancouver dataset was highly successfully aligned within the maximum accepted time lapse of 30 minutes with a precision of 94%. By contrast, the London dataset achieves a good overall performance by means of an accuracy of 87% at a significantly lower precision level of 47%. The results on the sub-event detection task are ambiguous. Overall, the AHC-based approach tends to detect a significantly larger number of sub-events than the XMeans-based approach. Nevertheless, both approaches result in high Random Index (RI) scores which reflects the purity of the detected clusters. While in general high RI scores may also be the result of strong over-segmentation, the number of detected clusters with our runs differ significantly.

Table 3: *MediaEval 2014 Benchmark* results for the synchronization task in terms of precision (P) and accuracy (A).

	Vancouver dataset		London dataset	
	P	A	P	A
AHC + MPEG7-SC	0.9412	0.7919	0.4722	0.8746
XMeans + JCD	0.5882	0.5701	0.3611	0.4676

Table 4: *MediaEval 2014 Benchmark* results for the sub-event detection task in terms of number of detected clusters (C), Random Index (RI), and F1-score (F1).

	Vancouver dataset			London dataset		
	C	RI	F1	C	RI	F1
run 1	379	0.9787	0.1012	368	0.9842	0.2614
run 2	709	0.9782	0.0505	709	0.9873	0.1687
run 3	91	0.9610	0.1087	91	0.9760	0.1331
run 4	81	0.9687	0.0890	81	0.9797	0.1653
run 5	98	0.9727	0.1079	98	0.9797	0.1653

4. CONCLUSION

In this paper we presented two multimodal approaches for the synchronization of multi-user galleries and for the detection of sub-events. The results obtained on the SEM datasets indicate the potential of the combination of visual and time information for the tasks. An open issue is the detection of sub-events that are visually highly similar and that take place in a short time period.

Acknowledgments

This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010, by the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/22573/33955, and by the iAD center for Research-based Innovation (project number 174867) funded by the Norwegian Research Council.

5. REFERENCES

[1] N. Conci, F. D. Natale, and V. Mezaris. Synchronization of multi-user event media (SEM) at MediaEval 2014: Task description, datasets, and evaluation. In *MediaEval 2014 Workshop*, 2014.

[2] M. Lux. Lire: Open source image retrieval in java. In *ACM Int. Conf. on Multimedia*, pages 843–846, 2013.

[3] M. Riegler, M. Lux, and C. Kofler. Frame the crowd: Global visual features labeling boosted with crowdsourcing information. In *MediaEval 2013 Workshop*, 2013.

[4] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.