Event Video Retrieval using Global and Local Descriptors in Visual Domain

Jennifer Roldan Carlos*, Mathias Lux*, Xavier Giro-i-Nieto†, Pia Munoz* and Nektarios Anagnostopoulos*
*Klagenfurt University
Klagenfurt, Austria
Emails: jroldancarl@gmail.com, mlux@itec.aau.at, piamunozt@gmail.com, nek.anag@gmail.com
†Universitat Politecnica de Catalunya
Barcelona, Catalonia/Spain
Email: xavier.giro@upc.edu

***Abstract*—With the advent of affordable multimedia smart phones, it has become common that people take videos when they are at events. The larger the event, the larger is the amount of videos taken there and also, the more videos get shared online. To search in this mass of videos is a challenging topic. In this paper we present and discuss a prototype software for searching in such videos. We focus only on visual information, and we report on experiments based on a research data set. With a small study we show that our prototype demonstrates promising results by identifying the same scene in different videos taken from different angles solely based on content based image retrieval.**

I. Introduction

Many people like to share their experiences with friends. A large part of them uses the internet to publish and send pictures and videos from what they have seen, visited and experienced. YouTube alone currently has more than 300 hours worth of videos uploaded every minute[1]. Especially for large events where lots of people attend, it is common to find multiple videos from the same time and same location on YouTube, Facebook and alike, and it is hard to keep track on which videos show what.

In this paper we present a prototype for near duplicate visual search in videos. With such a prototype one can search for visually similar video frames throughout a collection of videos and eventually find those that have been taken from the same scene. For input our system relies on a video frame or an image. With the given query the system finds videos, where similar frames occur, ranks them by the relevance of the frames and the amount of frames found, and returns a list of videos with the relevant frames highlighted (cp. Figures 1, 2). For indexing we sample equidistant frames and use both, global and local features, for search. Result aggregation is done by late fusion.

The overall goal of the prototype is to give a proof of concept that visual search can be used to identify videos from events, where multiple videos have been recorded from the same scene. Based on the visually similar frames we assume videos can be hyperlinked or even roughly synchronized. We show the applicability of our approach by using the Jiku Mobile data set [1], which features videos taken from different users from a set of events; including dancing, singing and sports. For each event, multiple temporally and spatially overlapping videos are available.

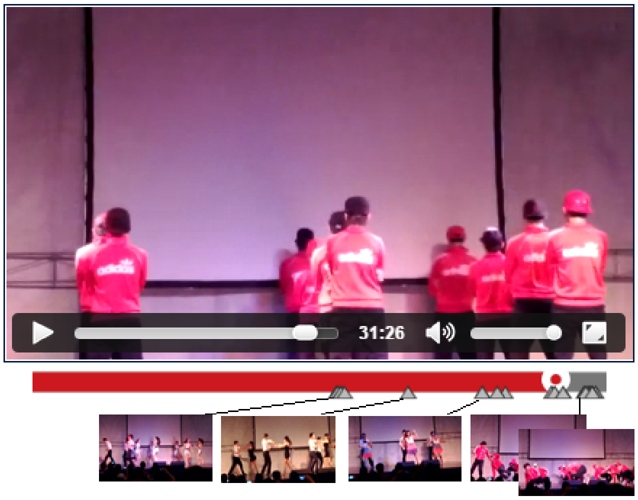


Fig. 1. Screen shot of a single result with the frames visually similar to the query highlighted.

A. Related Work

Related work in this field is, while it being an obvious approach, rather sparse. Most important related initiatives video and multimedia retrieval challenges such as TRECVID [2], MediaEval [3] and ImageCLEF [4]. While the tasks of the initiatives are changing, near duplicate frame search either has been a task or it has been used as means to an end for tackling one of the tasks. Our prototype is different to previous approaches as it incorporates the SIMPLE descriptors, which are local features used for the first time in the field of video retrieval.

In [5] the authors present a system also focusing on videos taken at events. However, they employ a more controlled and holistic approach. Videos recorded with their software are automatically enriched with meta data, ie. sensor readings, which allows for faster and easier retrieval, while we do

[1] https://www.youtube.com/yt/press/, visited 2015-03-03

not restrict the video recording procedure and operate on visual data only. In [6] the authors present a system, which automatically creates an event summary based on different videos from different users and view points. The system, called Jiku Director, operates on the same data set as our prototype does, but relies solely on meta data. The main contribution is the creation of the summary, not the retrieval of scenes.

A similar case using a large scale dataset is presented in [7]. The dataset contains of 3,800 hours of newscasts and features 200 queries for retrieval evaluation providing a ground truth. The queries are images and have to be found in the video streams, an approach the authors call image-to-video, *I2V*. Moreover, the authors present a system operating on the data set in [8].

II. Our prototype

In our demo application we focus on content based video indexing and retrieval to match example query content to target video content by extracting and indexing visual feature descriptors. Each descriptor can be considered as an *independent retrieval model* [9] which at some point needs to be fused. Mainly, two types of fusion schemes are considered. In *early fusion* the retrieval models are integrated from the start and afterwards a multimodal representation is learned. *Late fusion* approaches on the other hand infer similarity directly from unimodal features and integrate results at the end [10].

In the demo, we employ a late fusion model based on multiple global features using a single visual example. The goal of late fusion techniques is the combination and re-score or re-rank of the initial result lists into a single final list. Before fusing the top hits from different lists it is required to truncate to the top N results and normalize them either by rank

$$\bar{R}_k(n) = \frac{N + 1 - R_k(n)}{N}$$

or by score

$$\bar{R}_k(n) = \frac{R_k(n) - min(R_k)}{max(R_k) - min(R_k)}$$

where R_k is the initial result (rank or score) from the retrieval model k. For our demo we apply the sum approach, where either normalized ranks or scores are summed up (cp. fusion strategies in [11]):

$$R_t(n) = \sum_k (R_k(n)) = R_1(n) + R_2(n) + ... + R_K(n)$$

For late fusion we used three different global features, (i) *CEDD* [12], a compact joint histogram of fuzzy color and texture, (ii) the *auto color correlogram* [13], a color feature that measures how often a color encounters itself in a neighborhood, and (iii) the *pyramid histogram of oriented gradients* (PHOG) [14], a fuzzy gradient histogram organized in a spatial pyramid.

In addition to the global descriptors, we also introduce localized version of CEDD employing the SIMPLE model [15], which has outperformed classical local features in many scenarios. SIMPLE uses a key point detector to find salient points on different scales. Based on the scale the point has been found, a local image patch is indexed with a compact and composite descriptor. Following that, the *bag of visual words* model is used to aggregate local features into histograms. We used SIMPLE with the CEDD feature, the SURF key point detector [16], and k-means to create a visual vocabulary of 512 visual words. All of the features were extracted with the open source library LIRE [17].

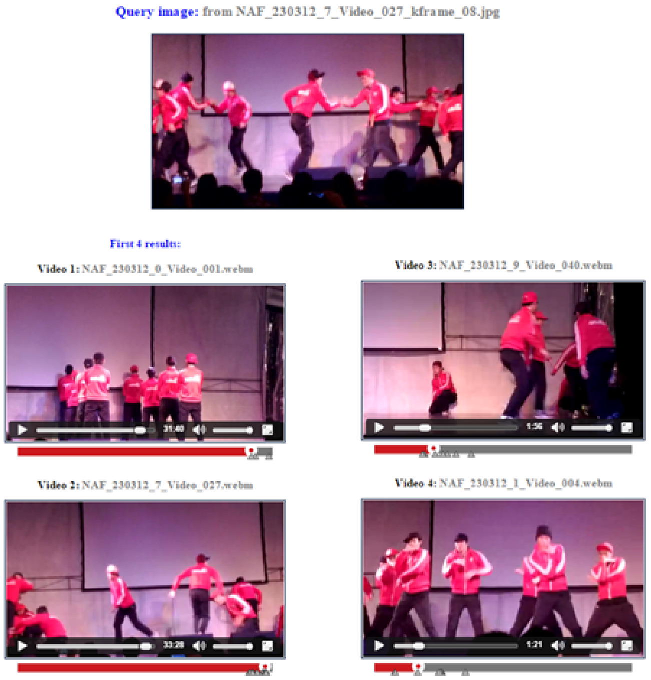

Query image: from NAF_230312_7_Video_027_kframe_08.jpg
First 4 results:
Video 1: NAF_230312_0_Video_001.webm
31:40
Video 3: NAF_230312_9_Video_040.webm
1:56
Video 2: NAF_230312_7_Video_027.webm
33:28
Video 4: NAF_230312_1_Video_004.webm
1:21

Fig. 2. Screen shot of the application showing a query and the first four results.

As search process is based on frames within the videos and the result list is also composed of video frames, our system aggregates the frames as a last step. Based on the top 40 hits for each query, we determine the four best matching videos and present them to the user while visualizing the time location where the matching frames have been actually found, as shown in Fig. 2. For this reason, the final ranked list of videos is based on their best matching frame, ie. the first frame defines the best matching video, the first frame of a different video in the result list of frames defines the second best matching video, etc.

III. Experiments

We used the Jiku Mobile data set [1] for our study, which is a set of 473 video clips taken at five different social events. The videos were recorded by different people from different angles. They feature pairwise overlap time- and scene-wise. For our experiments we indexed 356 randomly selected videos based on equidistant frames, using one frame per second. A set of 412 queries of different performances (cp. Figure 3) in

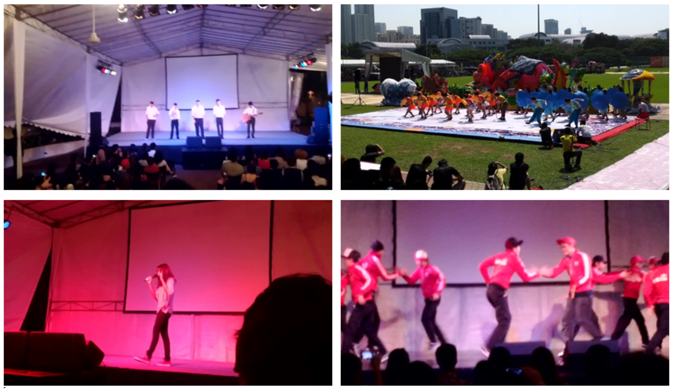


Fig. 3. Sample queries showing scenes from indoor and outdoor events as well as different points of view.

the social events was created manually. We aimed to cover different aspects, like for instance, outdoor, indoor scenes, colorful, and simple scenes.

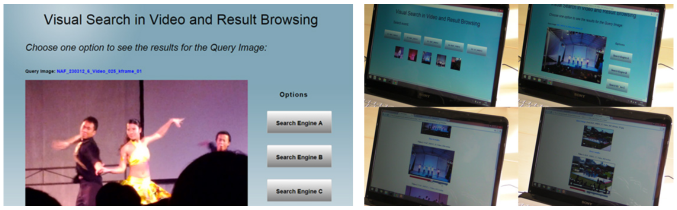
Based on the 412 queries we created a benchmarking data set. We tested if all the queries are to be found within the video data set. Our tests have shown that the video from which the query frame was extracted was ranked at the first position for 96% of the cases (cp. Table I). This confirms that the subsampling of one frame per second is enough for the data set to yield meaningful and accurate results with our approach.

TABLE I
RESULTS OF THE TESTS ON WHERE THAT ACTUAL VIDEO CAN BE FOUND IN THE RESULTS. THE FIRST TWO COLUMNS GIVE THE TWO DIFFERENT TESTED FEATURE FUSION APPROACHES, THE THIRD ONE GIVES THE RESULTS ON THE USE OF SIMPLE-CEDD.

	Sum of Ranks	**Sum of Scores**	**SIMPLE**
Precision @ 1	0.964	0.966	0.908
Precision @ 2	0.976	0.976	0.927
Precision @ 3	0.978	0.978	0.927
Precision @ 4	0.981	0.983	0.927

In order to test our prototype, we implemented a semi-interactive web based interface which allows users to dynamically select a query image and see the search results from three search configurations. In particular, the interface presents to the users a manually selected set of query frames from five social events of the Jiku Mobile data set. Users can explore the results from three configurations, named *search engines* for the sake of the test. These three approaches have been labeled as *search engine A* (for sum of ranks fusion of global features), *search engine B* (for sum of scores fusion of global features) and *search engine C* (for the use of SIMPLE based local features).

We asked the users to test which of the three search engines satisfied the users' needs, and which of them gives subjectively better results by mining ie. more accurate or broader. We did not want to give the users a goal beside explaining them what the data set and the queries meant. It was up to them to decide if the search engines returned what seemed *natural* to the users. Each user freely chose different queries and investigated the results provided by the three search engines. In that sense it was a heuristic evaluation asking experts on the overall performance. All six test subjects had been working in the field of computer science, and especially multimedia research for several years.



Visual Search in Video and Result Browsing

Choose one option to see the results for the Query Image:

Query Image: NAF_230312_6_Video_025_kframe_01

Fig. 4. Screen shot of the application showing the query interface (left) and a view on the test subjects environment (right).

For evaluating the results we employed a *thinking aloud test* setup as described in [18]. It consisted of two parts. (i) Part one is a hands-on experience by different participants who used our tool. In this part we asked the participants to voice their thoughts and we did not interfere or encourage them. The sessions were recorded with one video camera over their shoulders capturing the mechanical interaction with the tool. (ii) Part two is an open interview reflecting his experience with the tool. Users are asked during the interview what they think about the tool and which conclusions they extract from this test. Is it a useful tool? Does this tool cover their expectations? After the tests we reviewed and transcribed all the interviews and test sessions. Based on the transcripts and the notes taken we discussed the results and concluded on the test.

As a general overview, we noticed the users were expecting to see *visually* similar scenes or *the same performances* in the results of the search. They particularly looked out for hints that this is a video showing the very same event, and eventually the same part of the event. All of them appreciated the similarities in the background, the stage or the number of people which are shown in the results. However, the main expectation they had was to find the same performance from *different point of view*.

The participants choose the query image based on their intuition of what would result interesting, ie. they were driven by their own curiosity. They were driven by many reasons, as for example the simplicity of the scenario with specific people on it, or colorful scenario outdoor crowded of people. Other reasons are the out-of-the-ordinary background color or a specific performance with out-of-the-ordinary movements.

After choosing a query image, some users were expecting to see directly videos showing the performance. They realize later that the results are shown in the time line. They expressed their view about the time line as a great tool to use in the demo application, as it allows the user to go directly to the final results without the need to watch the whole video. To investigate subtle and non-obvious differences between the different search engine, participants opened multiple tabs in the web browser and compared the results by switching between them.

As an overall impression, for the search engines A and B, which are the sum of ranks and sum of scores fusion of global features, user comment they are good approaches for abstract exploratory search with a query as an example, and when searching for scenes with the same viewpoint of the stage, even with different sub-events. For search engine C, which is the SIMPLE based local features approach, all the users agree on this is the search engine that fits better when the user is searching for semantically similar content. Mostly, it shows the same performance with different viewpoints. Moreover, this search engine tends to retrieve fewer hits, which is (i) it is less confusing for the user and (ii) users need fewer steps to reach the right time point.

IV. Conclusion

In this paper we have presented a prototype implementation for video search based on frames and visual information retrieval. We further reported on tests using a freely available research data set. Our experiments have indicated that both methods employed, (i) late fusion of global features and (ii) use of SIMPLE based local features, have their merits for different types of queries in the investigated use case. Using SIMPLE search is more accurate and can retrieve the same scene from different angles, while global features present a broader picture, match scenes with similar background and allow for a more exploratory type of search.

However, for practical use of our method we have to take into account the amount of indexing time. Local features of course take additional time as a code book has to be created, so for practical use the code book should be pre-computed.

In the future we want to try our method, especially visual search in videos based on SIMPLE on larger data sets and we want to compare it to more traditional local feature approaches like SIFT/SURF BoVW [19]. We further aim to fuse local and global features, which may allow us to get best of both worlds, and, for practical use, we want to speed up indexing time.

Acknowledgements

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/25557/37319. It has also been developed in the framework of the project TEC2013-43935-R, financed by the Spanish Ministerio de Economa y Competitividad and the European Regional Development Fund (ERDF).

References

[1] M. Saini, S. P. Venkatagiri, W. T. Ooi, and M. C. Chan, "The jiku mobile video dataset," in *Proceedings of the 4th ACM Multimedia Systems Conference*, ser. MMSys '13. New York, NY, USA: ACM, 2013, pp. 108–113. [Online]. Available: http://doi.acm.org/10.1145/2483977.2483990

[2] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[3] M. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. Sutcliffe, J. Choi, and G. J. Jones, Eds., *MediaEval 2014 Multimedia Benchmark Workshop*, ser. CEUR Workshop Proceedings, vol. 1263, Oct. 2014.

[4] H. Müller, P. Clough, T. Deselaers, and B. Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[5] P. Seshadri, M. Chan, W. Ooi, and J. Chiam, "On demand retrieval of crowdsourced mobile video," *Sensors Journal, IEEE*, vol. PP, no. 99, pp. 1–1, 2014.

[6] D.-T.-D. Nguyen, M. Saini, V.-T. Nguyen, and W. T. Ooi, "Jiku director: A mobile video mashup system," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 477–478.

[7] A. Araujo, J. Chaves, D. Chen, R. Angst, and B. Girod, "Stanford I2V: A News Video Dataset for Query-by-Image Experiments," in *Proc. ACM Multimedia Systems*, 2015.

[8] A. Araujo, D. Chen, P. Vajda, and B. Girod, "Real-time query-by-image video search system," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 723–724. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654867

[9] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 172–179.

[10] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[11] K. Mc Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Image and video retrieval*. Springer, 2005, pp. 61–70.

[12] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Computer Vision Systems, 6th International Conference, ICVS 2008, Santorini, Greece, May 12-15, 2008, Proceedings*, 2008, pp. 312–322.

[13] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, 1997, pp. 762–768.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.

[15] C. Iakovidou, N. Anagnostopoulos, A. C. Kapoutsis, Y. Boutalis, and S. A. Chatzichristofis, "Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: the SIMPLE answer to effective content based image retrieval," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2014, pp. 1–6.

[16] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[17] M. Lux and S. A. Chatzichristofis, "Lire: lucene image retrieval: an extensible java CBIR library," in *Proceedings of the 16th International Conference on Multimedia 2008,*, Vancouver, Canada, Oct 2008, pp. 1085–1088.

[18] T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *Professional Communication, IEEE Transactions on*, vol. 43, no. 3, pp. 261–278, 2000.

[19] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.