Global vs. Local Feature in Video Summarization: Experimental Results.

Marian Kogler, Manfred del Fabro, Mathias Lux, Klaus Schöffmann, and Laszlo Böszörmenyi

Institute for Information Technology
Klagenfurt University
Universitätsstrasse 65–67
9020 Klagenfurt, Austria
{mkogler, manfred, mlux, ks, laszlo}@itec.uni-klu.ac.at

Abstract. We investigate the usefulness of local features in generating static video summaries. The proposed approach is based on bag of visual words using SIFT features. In an explorative experiment we compare this approach to summaries generated with the help of global features. As a resume we conclude that the local feature based approach does not outperform the other ones, however, it seems to be more stable.

1 Introduction

In the last decade the importance of videos conveying information has increased, which is accompanied by the need to store, organize and index the multimedia content appropriately, in order to support the user in retrieving videos. A lot of video clips are produced, broadcasted, shared and stored every day by professionals, amateurs and hobbyists. Finding videos matching the actual information need of a user proves to be a hard problem. *Video abstracts*, or video summaries, aim at presenting the semantics and content of a clip in minimized time and space to allow fast assessment of video clip relevance. In this paper we focus on static methods: still image summaries showing keyframes from the video.

Generally speaking a video abstract should maximize the (semantic) information transported by the summary while minimizing time and pixels needed to show, store and assess the summary. We have created a keyframe selection tool (discussed in detail in [13]), which implements summarization of video clips by keyframe extraction based on global image features.

We further extended the tool to support extraction based on local features in order to find out, whether the summarization process leads to a better summarization of video clips. We apply SIFT features proposed by Lowe [12] to extract feature vectors from salient keypoints of an image. The salient points and their 128 dimensional feature vectors are interpreted as local features describing a video frame. For pairwise comparison of frames we employ the *bag of visual words* approach (see e.g. [5], [10], [20], [18], [16]). All local features are clustered using K-Means [9]. The cluster centers are interpreted as reference feature for the

whole cluster and are called *visual word*. A single frame is then represented by a histogram, called *local feature histogram* [6], denoting the occurrence of visual words within the frame. Figure 1 depicts the described approach, in order to get a better understanding. For keyframe selection the local feature histograms are k-medoid clustered [7] and cluster medoids are selected as representative keyframes of a frame cluster. Cluster medoids are ranked based on the cluster size.

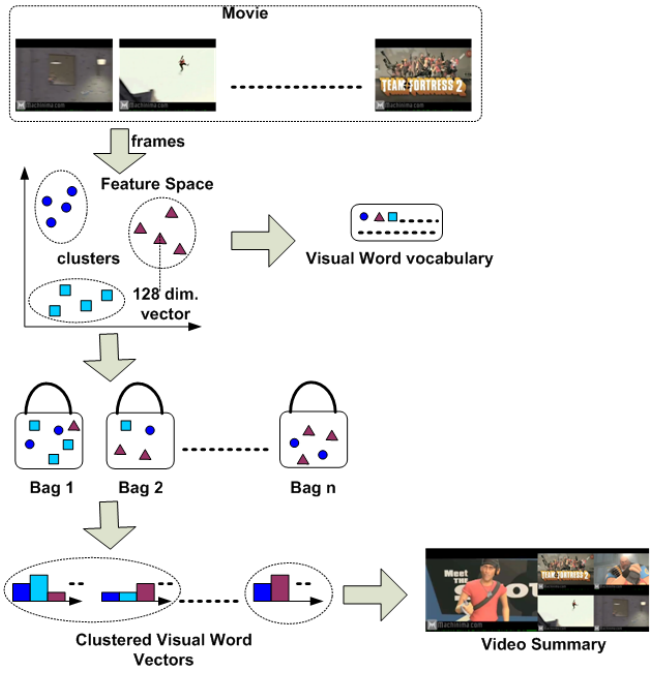

Fig. 1. Bag of Visual Words for Video Summarization

We apply the mentioned bag of visual words technique in our video clip summarization, in order to achieve a video summary more suitable for the user to assess the videos relevance. This should facilitate the search process of the user in a huge multimedia database, by depicting more meaningful images in a video summary.

The remainder of this paper is composed as follows: in Section two we give a short overview of video summary techniques. Section three covers our exploratory study which was conducted in order to test the performance of global vs. local features. Section four concludes our paper, discusses contributions and lessons learned and gives an outlook on future work.

2 Related Work

The main idea behind video summarization is to take the most representative and most interesting segments of a long video in order to concatenate it to a new, smaller, sequence. Video summarization has been investigated for several years already. Nevertheless, in a recent review [17] – that contains more than 160 references! – the authors conclude that "video abstraction is still largely in the research phase".

Proposed methods can be classified by the low-level features which are used for content analysis. In general, video summarization is either performed by image features (e.g. [2]), audio features (e.g. [19]), textual features (e.g. [4]), or a combination of several features (*multi-modal* methods, e.g. [14]).

A further classification can be performed on the presentation of the summary. *Static methods* use representative keyframes (e.g. in a storyboard visualization). *Dynamic methods* use video skims (e.g. a slide-show of keyframes). The static method has the advantage that a user can more quickly watch the entire summary, while the dynamic approach may allow a more comprehendible summary not only because usually audio playback is also available. In addition, *interactive video summarization* methods allow a user to selectively see parts of the summary according to a query.

Another classfication has been presented by Money et al.[15]. They classified video summarization methods into *internal*, *external*, and *hybrid* ones. The most common ones are internal methods, where content analysis is performed directly on the video stream. External methods (e.g. [21]) use information not directly contained in the video stream (e.g. manual annotation), and hybrid methods use a combination of both.

Recent efforts try to create personalized video summaries by integrating the users' interest. For instance, Matos et al.[14] use multimodal analysis together with a model of the *users' arousal*. Lie et al. [11] propose another personalized video summarization system. Their system allows a user to formulate preferences on semantic events like the appearance of humans, the happening of specific events (explosions, moving objects, zoom-in), and the differentiation between indoor and outdoor scenes. Another interesting approach has been presented by Bailer et al. [1]. They propose a collaborative summarization method, where several methods for content segmentation and segment selection are combined and finally fused together in order to produce the video summary.

3 User Study

We conducted an exploratory evaluation, where users had to choose their favorite summaries depicting the corresponding videos in best manner. We distinguished between four low level features (ACC [8], CEDD [3], RGB color histograms, SIFT), which led to four different summaries for each video clip. In a previous study [13] we investigated a number of global features. Summaries generated on the basis of the ACC, CEDD and RGB features were favored by the users and therefore selected in the actual study to compete with local SIFT features.

One summary consisted of five keyframes extracted by our tool. These keyframes were arranged in a single summary image which was presented to the user. We analyzed four short video clips ranging from news to animations. Because videos longer than five minutes probably cover too much information, which cannot be depicted properly in a video summary consisting of five still images, we only considered short ones. A further reason for selecting short clips is, that video clips recorded by users, in order to retain a moment of attraction, usually do not take longer than three minutes. This assumption is based on the observation that the average length of a video clip posted on YouTube is only 2 minutes and 46.17 seconds[1].

Table 1. videos used for exploratory study

Title	**Length**
iPhone commercial[2]	76 s
dinosaurs vault[3]	48 s
hurricane IKE - news reporter almost washed away[4]	30 s
shrek[5]	48 s

Each video is summarized by a full sized frame of the biggest cluster (the cluster with most frames) on the left, followed by four frames half in width and height on the right representing smaller clusters. Figure 2 shows a sample visualization of a video summary created by our tool for the Shrek video.

Fig. 2. Visualization of our video summary (based on CEDD) depicting a video clip

Each participating user had to assess four summaries (4 points for the best, down to 1 point for the worst) for each video clip, which led to a total of 16

[1] Statistics from http://ksudigg.wetpaint.com/page/YouTube+Statistics
[2] http://www.youtube.com/watch?v=2k3zvI2tyPM
[3] http://www.youtube.com/watch?v=Dim0INyvJdw
[4] http://www.youtube.com/watch?v=SYl9mgFhe2o
[5] http://www.youtube.com/watch?v=uvyelwDA0Ws
(last checked: 2009-09-22)

summaries. The user group consisted of 9 people (5 female and 4 male); ages ranging from 20 to 30 years.

3.1 Results

There was no clear winner in our experiment. All four selected image features got similar ratings from the test persons as Figure 3 shows. The summaries based on CEDD have reached the highest score (104 points), followed by our SIFT based visual bag of words approach (91 points), ACC (87 points) and the color histogram (78 points). The scores for each single video are shown in Table 2.

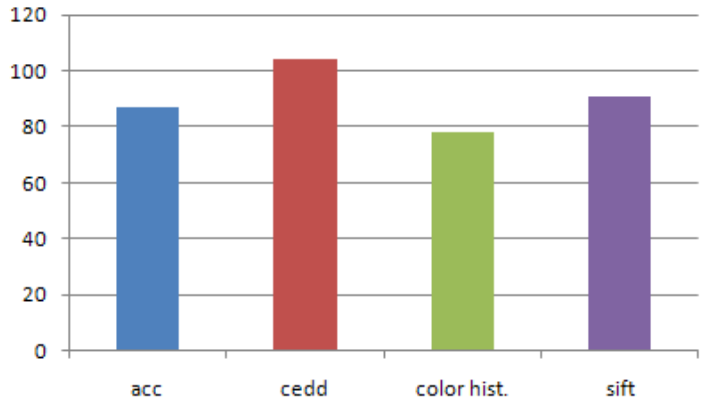

Fig. 3. Summed user ratings of the low-level features used for keyframe selection

Table 2. Rating of the features for each video

	ACC	CEDD	color histogram	SIFT
iPhone	29	23	17	21
News	20	31	16	23
Shrek	11	33	25	21
Dinosaur	27	17	20	26

It can be seen that the type of the chosen features heavily depends on the type of the video. While CEDD produces very good results for the news video and the clip of the movie Shrek, it performs rather poor for the animation with the dinosaurs. On the other side, ACC reaches a high score for the iPhone commercial and the dinosaurs animation, but it is a bad choice for the Shrek clip. Our SIFT-based bag of visual words approach never reached the best score, but also never

performed worst, which can be seen easily in Figure 4. In three cases (iPhone, news and dinosaurs) it reached the second highest score and in the fourth case (Shrek) it reached the third place. Therefore, it seems that this approach based on local image features produces more stable results than the ones based on global image features. This assumption is also supported by the deviation of the samples, given in Table 3. The local feature approach in our experiments features lowest standard deviation (SIFT, 0.84) and can be considered the most stable approach for our test set.

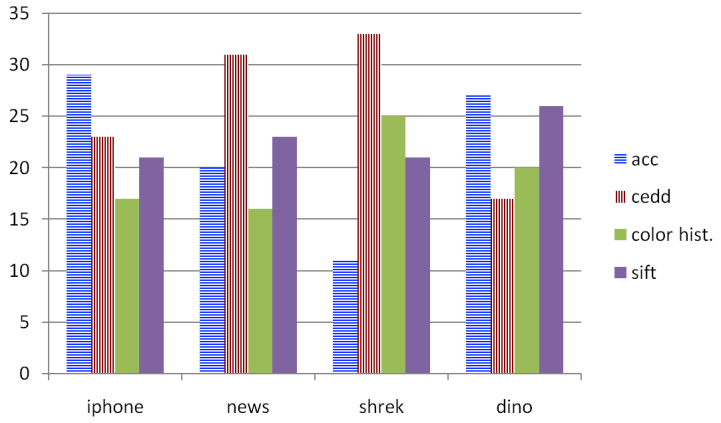

Fig. 4. User ratings of the low-level features used for keyframe selection per video

Table 3. Standard deviation for the ratings of the selected visual features

Feature	Standard deviation
ACC	1.18
CEDD	1.14
color histogram	1.21
SIFT	0.84

4 Conclusion

We presented the results of a study, where users weighted the appropriateness of video summaries based on their ability to describe a short video clip. Main focus of our investigation was the question whether local features achieve better summaries than global features. We employed a visual bag of words approach using

the SIFT descriptor and tested the approach on 4 videos with an exploratory study with 9 users. Results showed that while the local feature approach could not outperform the global feature approaches, it provides for our test data set and the test population the most stable results being ranked second three times and third one time. Even though the test data set and the population of the survey are too small to provide significant results, they allow the hypothesis that local features provide more stability than global features in general use cases and encourage further research on this.

With our implementation we could also see the difference in runtime between the different approaches. Extraction of SIFT features and finding of the visual words took ten times longer than the extraction of global features, say CEDD (70 vs. 700 ms per frame on a 2 GHz dual CPU workstation). While this can be further reduced using faster and optimized local features the whole process of extraction, clustering of the salient points and creation of the visual word vocabulary is significantly slower than a global feature based approach.

In the near future we will test the local feature approach in different domains including medical videos and user captured single shot videos. We hope that we gain insights on the applicability of local feature histograms and the overall performance in and throughout different domains.

5 Acknowledgments

This work was supported by the Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant 20214/17097/24774.

References

1. W. Bailer, E. Dumont, S. Essid, and B. Merialdo, *A collaborative approach to automatic rushes video summarization*, 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, 2008, pp. 29–32.
2. D. Borth, A. Ulges, C. Schulze, and T.M. Breuel, *Keyframe Extraction for Video Tagging and Summarization*, Proc. Informatiktage, 2008, pp. 45–48.
3. Savvas A. Chatzichristofis, Yiannis S. Boutalis, and Mathias Lux, *Selection of the proper compact composite descriptor for improving content based image retrieval*, The Sixth IASTED International Conference on Signal Processing, Pattern Recognition and Applications SPPRA 2009, 2009.
4. G. Ciocca and R. Schettini, *An innovative algorithm for key frame extraction in video summarization*, Journal of Real-Time Image Processing **1** (2006), no. 1, 69–88.
5. T. Deselaers, L. Pimenidis, and H. Ney, *Bag-of-visual-words models for adult image classification and filtering*, Proc. 19th International Conference on Pattern Recognition ICPR 2008, December 8–11, 2008, pp. 1–4.
6. Thomas Deselaers, Daniel Keysers, and Hermann Ney, *Features for image retrieval: an experimental comparison*, Inf. Retr. **11** (2008), no. 2, 77–107.

7. Youssef Hadi, Fedwa Essannouni, and Rachid Oulad Haj Thami, *Video summarization by k-medoid clustering*, SAC '06: Proceedings of the 2006 ACM symposium on Applied computing (New York, NY, USA), ACM, 2006, pp. 1400–1401.
8. Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih, *Image indexing using color correlograms*, CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97) (Washington, DC, USA), IEEE Computer Society, 1997, p. 762.
9. A. K. Jain, M. N. Murty, and P. J. Flynn, *Data clustering: a review*, ACM Comput. Surv. **31** (1999), no. 3, 264–323.
10. Yu-Gang Jiang and Chong-Wah Ngo, *Bag-of-visual-words expansion using visual relatedness for video indexing*, SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM, 2008, pp. 769–770.
11. W.N. Lie and K.C. Hsu, *Video summarization based on semantic feature analysis and user preference*, Proc. IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing 2008, 2008, pp. 486–491.
12. David G. Lowe, *Object recognition from local scale-invariant features*, ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2 (Washington, DC, USA), IEEE Computer Society, 1999, p. 1150.
13. Lux, M.; Schöffmann, K.; Marques, O.; Böszörmenyi, L., *A novel tool for quick video summarization using keyframe extraction techniques*, Proceedings of 9th Workshop on Multimedia Metadata(WMM'09), CEUR Workship Proceedings, vol. 441, march 19–20 2009.
14. N. Matos and F. Pereira, *Using MPEG-7 for Generic Audiovisual Content Automatic Summarization*, Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on, 2008, pp. 41–45.
15. A.G. Money and H. Agius, *Video summarisation: A conceptual framework and survey of the state of the art*, Journal of Visual Communication and Image Representation **19** (2008), no. 2, 121–143.
16. C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, B. Huurnink, J.C. van Gemert, J.R.R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M. van Liempt, R. van Balen, F. Yan, M.A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.M. Geusebroek, T. Gevers, M. Worring, A.W.M. Smeulders, and D.C. Koelma, *The mediamill trecvid 2008 semantic video search engine*, (2009).
17. B.T. Truong and S. Venkatesh, *Video abstraction: A systematic review and classification*, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) **3** (2007), no. 1.
18. J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, *Real-time bag of words, approximately*, ACM International Conference on Image and Video Retrieval, 2009.
19. M. Xu, NC Maddage, C. Xu, M. Kankanhalli, and Q. Tian, *Creating audio keywords for event detection in soccer video*, Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on, vol. 2, 2003.
20. Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo, *Evaluating bag-of-visual-words representations in scene classification*, MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval (New York, NY, USA), ACM, 2007, pp. 197–206.
21. D. Zhang and S.F. Chang, *Event detection in baseball video using superimposed caption recognition*, Proceedings of the tenth ACM international conference on Multimedia, ACM New York, NY, USA, 2002, pp. 315–318.