

# Video Content Representation Using Recurring Regions Detection

Lukas Diem<sup>1</sup>(✉) and Maia Zaharieva<sup>1,2</sup>

<sup>1</sup> Multimedia Information Systems Group, University of Vienna, Vienna, Austria  
diem@cs.univie.ac.at

<sup>2</sup> Interactive Media Systems Group, Vienna University of Technology,  
Vienna, Austria

**Abstract.** In this work we present an approach for video content representation based on the detection of recurring visual elements or regions. We hypothesize that such elements play a potentially central role in the underlying video sequence. The approach makes use of fundamental intrinsic properties of a video and, thus, it does not make any assumptions about the video content itself. Furthermore, our approach does not require for any training or prior knowledge about the general settings and video domain. Preliminary experiments with a small and heterogeneous dataset of web videos demonstrate the potential of the approach to be employed as a compact summary of the video content with focus on its central visual elements. Additionally, resulting representations enable the retrieval of video sequences sharing common visual elements.

**Keywords:** Video content-based analysis · Recurring regions · Video representation

## 1 Introduction

Video content representation plays a crucial role for consumers in assessing its relevance to the personal interests and needs. Automated generation of content representation is a non-trivial task. The core challenge considers the selection of central and relevant aspects of the underlying content while preserving simplicity and low redundancy.

The assessment of relevance and significance is a high-level task that usually requires for additional knowledge about the content or its settings. For example, face detection can help to identify the main characters in a movie and to provide a compact overview of the playing actors. However, such an overview only presents a single aspect of the video, which does not allow for the assessment of the content itself. In a general setting, where there is no any prior knowledge about the explored video, a common way of content representation is by means of keyframes. The selection of keyframes is usually based on uniform sampling or on some visual criteria, e.g. large motion differences between consecutive frames often indicate substantial content change. However, this may still result in a large

amount of both irrelevant and redundant information for the video consumers that is not directly interpretable.

In this work we make use of a very fundamental characteristic of human perception. Humans tend to memorize things they are repeatedly confronted with. We, therefore, hypothesize, that elements recurring throughout a video sequence bear potential importance for the video content. Such elements can be characters or objects playing a central role. We propose an approach for video content representation based on the detection of recurring regions. The proposed approach makes use of the intrinsic properties of a video in terms of visual and motion coherence between consecutive frames. Therefore, we define a region as a visually and/or motion coherent element and, this, detected regions can represent a character (usually the main actors), an object, or a part of it. As a result, the approach does not require for any training in order to detect potential objects. Furthermore, it does not make any assumptions about the general settings and video domain (e.g. that objects of interest should be moving). The resulting video representation can be employed as a compact and interpretable video overview with the focus on central recurring elements. Furthermore, detected regions can be seen as an analogy to detected terms in a conventional text document. Thus, recurring regions enable novel, higher level similarity measure (e.g. in terms of term frequency - inverse document frequency) and retrieval approaches such as the search for videos sharing common visual elements.

In previous work, we transformed recurring element detection into the problem of matching and grouping local image descriptors [24]. In order to reduce complexity, we recently presented an approach that employs region segmentation and tracking [4]. In this paper, we extend our previous work on recurring region detection and introduce improvements of the underlying tracking and matching system.

This paper is organized as follows. Section 2 outlines related work in the context of video representations. We describe the proposed approach for the detection of recurring regions in a video sequence in Sect. 3. Section 4 presents the performed quantitative and qualitative analysis. Section 5 concludes the paper and provides an outlook for future work.

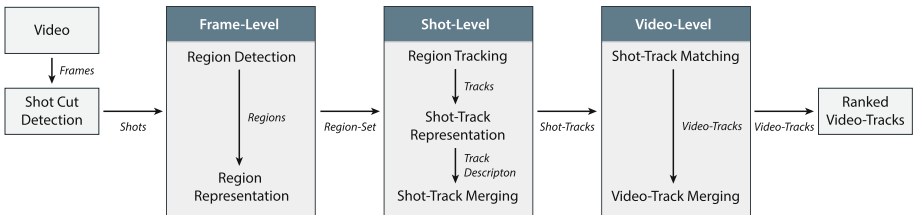
## 2 Related Work

Related work on video representations can be categorized into *visual-based*, *feature-based*, and *object-based* approaches. *Visual-based* summarizations provide an overview of the video content for the end-users. A commonly employed approach considers the use of key frames, e.g. by means of clustering all frames and selecting one or more frames per cluster for the final visual summary [1, 15, 21]. The number of key frames is a crucial parameter in this context since it directly influences the quality of the representation. Additionally, such visual summaries may contain several key frames depicting the same recurring elements in different environments that could be combined in a single representation. Another approach for video summarization visualizes a video as a graph. For example,

Zhang et al. group video shots into scenes and build a graph that models the scene structure of the underlying video [26]. A drawback of this representation is that a user does not get a visual summarization of the content (such as key frames) but only a set of nodes representing the detected video scenes. In contrast, Huang et al. depict a whole video collection as a graph of key frames and model the relationships between video shots [11]. Thus, the user is able to browse through the video collection and to retrieve videos containing similar shots.

*Feature-based* video representations are commonly employed for tasks such as near duplicate detection [13, 14], video retrieval, and activity recognition [20, 22]. The goal of a feature-based representation is to abstract the video content for further analysis rather than for visual summarization. Usually, frame- and shot-based features are aggregated to get a more compact video representation, e.g. by means of bag-of-words sampling and feature hashing [13, 14, 22]. Due to their nature extracted features are not directly interpretable by non-experts.

*Object-based* approaches aim at the detection and tracking of objects of interest within a video. This category can be further divided into supervised and unsupervised methods. Supervised object detection is commonly employed when the object of interest is known in advance, e.g. for surveillance tasks. In contrast, video segmentation methods usually have little assumptions on object categories. For example, several approaches employ long range point trajectories in combination with different grouping strategies (e.g. motion-based grouping) to segment video objects [6, 17, 18]. These methods report state-of-the-art results on segmentation benchmarks such as the VSB100 [8] and or FBMS [17]. A common limitation of such approaches is the assumption that objects of interest are moving within the video. Other methods employ a segmentation as graph-based grouping by building spatio-temporal graph representations [7, 10, 23]. Such approaches are often computationally complex since the graphs for a video representation easily become demanding in terms of space and computational power. Some recent approaches employ multiple segmentation proposals per frame that are temporally linked by similarity measures within a video sequence [2, 5, 12]. In general, the selection of the right hierarchy layer is highly depending to the video content and, thus, not feasible in a generic application.



**Fig. 1.** Overview of the proposed approach for recurring region detection.

### 3 Recurring Region Detection

Figure 1 outlines the proposed approach for the detection of recurring regions in a video sequence. We start with the detection of shot boundaries employing the method presented by Zeppelzauer et al. [25]. Within each shot, we detect and track regions using simple and efficient image-based techniques. Resulting shot-tracks are matched across the video in order to identify regions that recur repeatedly throughout the whole video sequence. The resulting video-tracks are sorted by their visibility score and represent the final set of detected recurring regions.

#### 3.1 Region Detection and Representation

We employ Statistical Region Merging (SRM) [16] to detect regions in each frame of the video. Core advantage of SRM is that it results in near object level segmentation. We describe each region by a color histogram with 72 bins (12 for achromatic colors, 12 for hue, and 5 for saturation) extracted from the HSV color space [4, 19]. Additionally, since very small regions are difficult to interpret and track over time, we merge regions with an area below a predefined threshold with the nearest neighboring region in terms of visual similarity.

#### 3.2 Tracking, Shot-Track Representation, and Merging

In order to track detected regions within each shot, we first initialize a *shot-track pool* with the region representations of the first frame in a shot. Following, we process all subsequent frames of the shot iteratively. For each region of a frame we consider all shot-tracks as possible assignments which are (1) within a predefined radius  $r$  of the current region position and (2) visually similar by means of the  $\chi^2$ -distance between the corresponding regions' descriptors,  $d_{\chi^2}$ . For all regions within the radius and below the maximum descriptor distance we compute a similarity score that penalizes distant assignments:

$$S_m = (1 - d_{\chi^2}) \cdot f\left(\frac{d_c}{r}, \mu, \sigma\right), \quad (1)$$

where  $f$  is a Gaussian weighting function,  $d_c$  the distance between the regions' centroids,  $r$  the predefined radius, and  $\mu$  and  $\sigma$  the parameters of the Gaussian distribution. The resulting list of potential assignments between previously tracked regions and the regions of the current frame is processed in decreasing order according to the achieved similarity score. Processed regions and shot-tracks are removed from the list to avoid multiple assignments. Eventually, for all regions, which are not assigned to an existing shot-track, we initialize a new shot-track and proceed with the next frame until all frames are processed.

Tracked regions (shot-tracks) are represented using the Color and Edge Directivity Descriptor (CEDD) [3]. Additionally, we employ the medoid region (i.e. the region with the minimum distance to all other regions within the track) as

a representative for the shot track. The representative region reduces the complexity of the following steps and is further employed for visualization purposes.

Objects within a shot are often represented by several shot-tracks due to occlusions and color variations within the object (e.g. a head is commonly split into a hair region and a face region). To account for such splits we introduce two refinement steps. First, we merge shot-tracks with similar medoid regions if the underlying tracked regions are immediate neighbors in the corresponding frames. Second, we additionally merge neighboring tracks if they carry similar motion patterns. Since the motion vectors defined by the region centroids of a shot-track are partly strongly jittering due to variations of the segmentation, we smooth the motion trajectories using an optimized version of a penalized least squares regression for discrete data [9]. To avoid merging stationary tracks we estimate the average motion per shot-track and allow for merges with a significant motion in comparison to the average motion of all shot-tracks of a shot. The average shot-track motion is estimated by

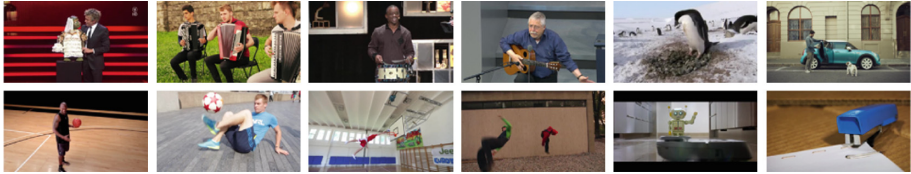
$$\bar{m} = \frac{1}{N-1} \sum_{t=1}^{N-1} \|c_{t+1} - c_t\|_2, \quad (2)$$

where  $N$  is the number of region centroids,  $\|\cdot\|_2$  is the Euclidean distance, and  $c_t$  is the smoothed region centroid at time  $t$ . For all neighboring shot-tracks, we compute the motion distance as the median of the (normalized) cosine distance between the motion vectors of consecutive frames and we merge shot-tracks with a distances below a predefined threshold. As a result, tracks that belong to the same rigid object but with a different visual appearance are merged.

### 3.3 Shot-Track Matching and Video-Track Merging

Recurring elements of a video usually appear in several shots. Therefore, we match detected shot-tracks across different video shots. The matching approach is split into two steps. First, we match shot-tracks similarly to tracking regions. We start by setting all shot-tracks of the first shot as initial video-tracks. For the following shot, we compute the  $\chi^2$ -distance between the medoid region of the video-tracks and the current shot-tracks and accept a match if the distance is below a predefined matching threshold  $t_m$ . Accepted matches are optionally refined to reduce the number of false positives. The refinement step considers the compactness of a shot-track before and after the merging step. We measure compactness in terms of average distance between all region descriptions of a shot-track. If the compactness changes significantly, i.e. the difference exceeds a predefined threshold  $t_c$ , the match is disregarded as it results in a visually inhomogeneous shot-track. If a shot-track is assigned to an existing video-track, the representative region of the track is updated otherwise a new video-track is initialized and we proceed with the next shot.

The final refinement step merges video-tracks that are visually similar and that fulfill one of the following requirements: (1) regions of the video-tracks are



**Fig. 2.** An overview of the employed dataset (each video is represented by a keyframe).

neighbors in all frames in which the video-tracks are visible, or (2) the video-tracks are never visible in a frame together. These requirements prevent merges of video-tracks that are spatially discontinuous.

Eventually, the final visibility score for each video-track corresponds to the fraction of time the video-track is visible (tracked) in the video sequence (e.g. a score of 0.75 means that a recurring element is visible in 75 % of all frames).

## 4 Evaluation

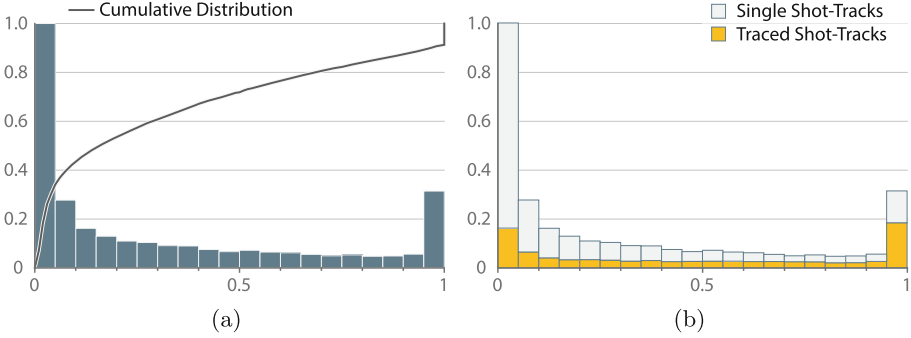
In this section we present the results of quantitative and qualitative experiments performed to evaluate the proposed approach.

### 4.1 Dataset

We employ the same dataset as presented in [4]. It consists of 12 YouTube videos with a duration between 1 and 6 min and an average duration of approx. 3 min ( $\sigma = 1.28$ ). Each video has 53 shots on average ( $\sigma = 29.8$ ) and the shots consist of 107 frames on average ( $\sigma = 44.3$ ). The genres of the videos are strongly varying, e.g. advertisement, music, sport (see Fig. 2). Additionally, all videos contain both static and moving cameras and objects.

### 4.2 Quantitative Evaluation

We first investigate the length of time a region is successfully traced within a shot, which corresponds to the visibility score of shot-tracks. Figure 3a visualizes the distribution of the visibility scores of the shot-tracks for all videos of the employed dataset. The distribution shows two major peaks, one for regions that are only visible in a few frames (visibility score below 0.05) and one for regions that are traced in nearly all frames of a shot (visibility score above 0.95). Overall, about 44 % of the regions are traced in less than 10 % of the frames. Such short shot-tracks often emerge from segmentation inaccuracies (e.g. region splits), or they represent visual elements (and potentially objects) which are only visible for a short period of time during the shot (e.g. due to camera or object motion). Additionally, about 33 % of the regions are traced and, thus, visible in more than 40 % of the shot. We assume that such longer lasting regions are the primary candidates for long-term recurring regions that can be employed to represent the content of the underlying video sequence. Therefore, we explore the traceability

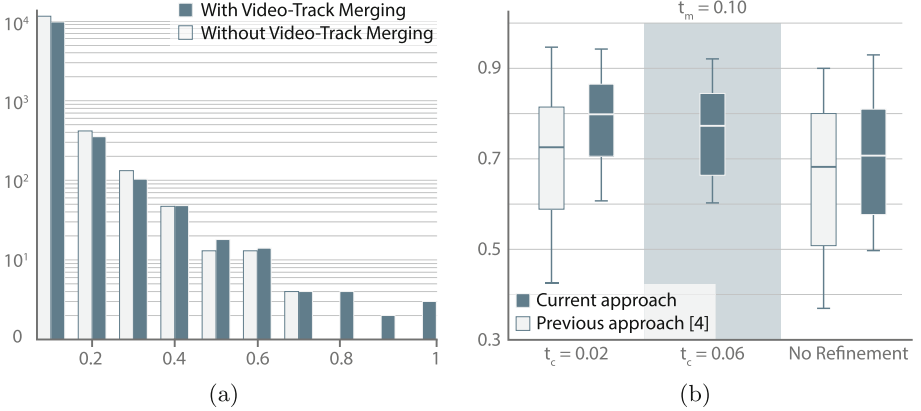


**Fig. 3.** (a) Distribution of the visibility scores of detected shot-tracks for all videos. (b) Distribution of the visibility scores of detected shot-tracks with respect to their traceability across different shots.

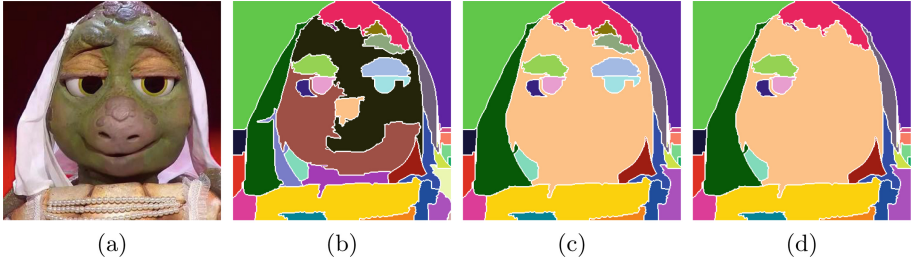
of shot-tracks with respect to their visibility score. Figure 3b shows the results of this evaluation. Long shot-tracks are commonly traced across several shots. This implies that visual elements that last in time in one shot will most probably reappear in the video sequence. On the opposite, short shot-tracks tend to remain unmatched throughout the underlying video sequence. Thus, such short shot-tracks can be considered for removal in order to improve the overall efficiency of the approach.

Next, we evaluate the influence of the proposed video-track merging step on the visibility scores of the complete video-tracks. As discussed in Sect. 3.3, we merge two visually similar video-tracks if either the underlying regions are neighbors in all frames they coincide or the underlying regions temporally complement each other. Figure 4a shows the distribution of the visibility scores of video-tracks with and without considering the merging step. The introduction of the step results in a merge of 21 % of all video-tracks. As a result, the visibility scores of the merged video-tracks increases, which is indicated by the shift of their distribution to the right. Overall, approximately 95 % of all video-tracks are visible in less than 10 % of the total video length. This confirms our assumption that there are only few central visual elements that recur throughout a complete video sequence. The quality of the detected regions in terms of content representation is discussed in Sect. 4.3.

Eventually, we compare the precision of tracing (matching) shot-tracks across different video shots with the approach presented in [4]. In this experiment, we match the shot-tracks of one shot with the shot-tracks of all other shots of the same video. All resulting matches are manually evaluated as true positives and false positives. Figure 4b shows the results for both approaches and for different parameter settings of the optional refinement step considering the compactness of shot-tracks:  $t_c = \{0.02, 0.06, \text{and } \textit{No Refinement}\}$ . Note that, the parameter  $t_c = 0.06$  is not evaluated by the previous approach. The average precision



**Fig. 4.** (a) Distribution of the visibility scores of the video-tracks for all videos. (b) Precision of the matching performance of shot-tracks for different parameter settings.



**Fig. 5.** Example shot-tracks: (a) (a detail of an) input frame, (b) tracked regions, (c) shot-tracks after merging by similarity, and (d) shot-tracks after merging by motion. Each shot-track is visualized by a pseudo-color (Color figure online).

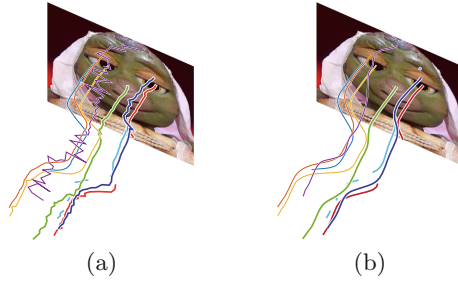
increases by more than 7% for the best parameter settings ( $t_m = 0.10, t_c = 0.02$ ) as a result of the improved region tracking and representation.

### 4.3 Qualitative Evaluation

We fix the matching parameters to  $t_m = 0.10$  and  $t_c = 0.06$  for all experiments. All examples are based on a video that shows the comedian Sascha Grammel talking to his turtle puppet Josie about marriage. The setting of the video is a gala night and the comedian is performing in front of the audience on a stage. The shots of the video show the puppet and the comedian from different perspectives as well as the audience.

We first analyze the quality of the proposed steps for merging shot-tracks (merging by similarity and merging by motion) and their impact on the final shot-tracks. Figure 5 shows detected regions in a frame and their alteration as a results of the merging steps. Originally, the face of the turtle is represented by

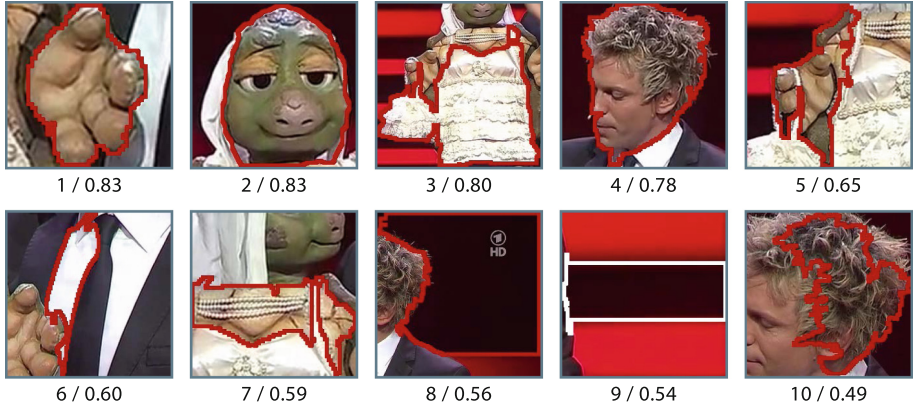




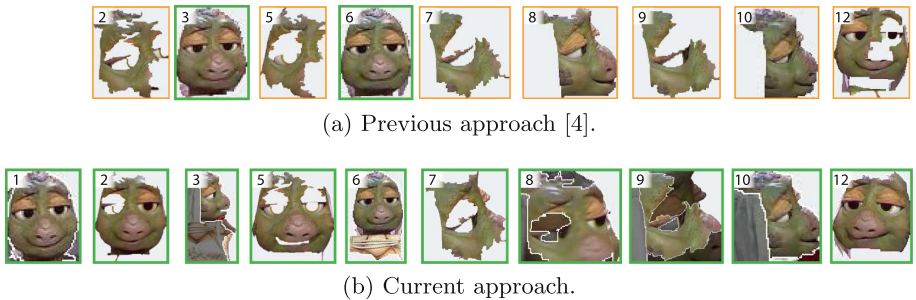
**Fig. 6.** Trajectories of the shot-tracks (a) before and (b) after smoothing. The thicker trajectories on the right in each image are successfully merged by the motion-based merging step.

several regions (and in following by several shot-tracks) due to the underlying over-segmentation. The first merging step, *merging by similarity*, is able to link the upper and lower parts of the face into a single shot-track (see Fig. 5c). However, the eyes and the top part of the face fail to merge since the descriptors of these shot-tracks achieve a low similarity score. Figure 5d shows the effect of the second merging step, *merging by motion*. The shot-tracks representing the right eye and eyebrow are merged into the shot-track representing the face. Figure 6 visualizes the trajectories of the tracked regions. Figure 6a demonstrates the necessity for smoothing trajectories in order to account for jittering regions. In contrast, smoothed trajectories in Fig. 6b reveal the underlying motion similarities between the shot-tracks. Since we employ a highly restrictive similarity estimation between trajectories to account for the heterogeneity of the underlying data, only a subset of the potential tracks is merged (cp. Fig. 5d).

Figure 7 shows the top 10 detected recurring regions for the investigated video sequence. The regions are sorted by the corresponding visibility score. This top 10 list demonstrates different aspects of our approach. Most of the top regions (8 out of 10) are part of the recurring regions identified by a human observer: the turtle puppet and the comedian. The regions, that belong to the turtle, are: 1 - left hand, 2 - turtle's face, 3 - wedding dress, 5 - right hand, and 7 - necklace. Regions, that belong to the comedian, are: 4 - head, 6 - shirt, and 10 - detail of the hair. The regions 8 and 9 show background elements. Since we only rank the regions according to their visibility, we cannot avoid that there are background regions within the top set of video-tracks. To remove such tracks from the results, additional information about the objects of interest is needed, which contradicts to the unsupervised methodology. Additionally, recurring background parts can be of interest during video analysis, for example when searching for videos sharing the same environment. We only visualize the top 10 regions since this limited amount of regions is already meaningful for the investigated video. However, a dynamic threshold can be employed to estimate the final number of regions from the visibility scores.



**Fig. 7.** Top 10 detected recurring regions for the *Sascha Grammel* video. Labels below the regions indicate the corresponding rank/visibility score.



**Fig. 8.** Representative regions of all shot-tracks of the turtles' face. The shot number is noted in the left upper corner. Orange (thin) and green (thick) borders indicate missed and successfully linked regions respectively (Color figure online).

Eventually, we investigate the shot-tracks that represent the face of the turtle, since this region lead to problems in the previous approach. Despite significant visual differences, we are able to match the turtle's face in all shots it is present. Figure 8 compares the results of both approaches. The green (thick) borders mark representative regions of the shot-tracks that are successfully matched into the same video-track. The representative regions of shots 2, 5, and 12 better represent the face due to the improved shot-track merging steps. The shot-tracks 7 and 9 of the current approach are part of a separate video track before the final video-track merging step. In this example the additional merging step worked as intended and significantly improved the result.

## 5 Conclusion

In this paper we presented a generic approach for the detection of recurring regions in a video sequence. This work is motivated by the fact that humans

usually memorize elements that are temporally persistent. As a result, recurrence in a video sequence can be seen as an indicator for importance. The result of our approach is a ranked list of detected recurring regions, which represents the content of the video sequence in a compact way. In addition to the visual representation, detected regions can be further employed for the retrieval of videos sharing common visual elements.

Performed experiments demonstrate both the potentials and the limitations of the proposed approach. *First*, achieved results show that detected recurring regions effectively capture visual elements that mostly play a central role for the underlying video sequence. However, a few recurring regions tend to capture background elements if the video sequence is shot in a single environment. The main reason is the aim of the approach to account for an arbitrary video data. As a result, both stationary and moving elements may capture a potential object of interest. However, the differentiation between a stationary object and the background is not feasible without any additional assumptions or a prior knowledge about the video settings. Nevertheless, background elements may still hold valuable information since they would support, for example, the search for video sequences shot in the same environment. *Second*, currently, detected recurring regions may still represent different parts of the same object. This is mainly due to the simplicity of the employed visual features. While, in general, they are more efficient than potential local features, they have lower descriptiveness. Nevertheless, the employed visual descriptors (and matching process) demonstrate high robustness across strongly varying visual appearances (cp. Fig. 8). A potential approach to improve the final visual representation is to consider the use of local features as final step to link recurring regions depicting different parts of the same object. Eventually, in future work, we will employ the proposed video content representation for the retrieval of videos sharing common visual elements. A core challenge in this context is the acquisition and annotation of a dataset of publicly available videos.

**Acknowledgment.** This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010.

## References

1. de Avila, S.E.F., Lopes, A.P.B., da Luz, A., de Albuquerque Araújo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.* **32**(1), 56–68 (2011)
2. Banica, D., Agape, A., Ion, A., Sminchisescu, C.: Video object segmentation by salient segment chain composition. In: *IEEE International Conference on Computer Vision Workshops*, pp. 283–290 (2013)
3. Chatzichristofis, S.A., Boutalis, Y.S.: CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) *ICVS 2008. LNCS*, vol. 5008, pp. 312–322. Springer, Heidelberg (2008)

4. Diem, L., Zaharieva, M.: Interpretable video representation. In: International Workshop on Content-based Multimedia Indexing, pp. 1–6 (2015)
5. Fragkiadaki, K., Arbelaez, P., Felsen, P., Malik, J.: Learning to segment moving objects in videos. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
6. Fragkiadaki, K., Zhang, G., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1853 (2012)
7. Galasso, F., Keuper, M., Brox, T., Schiele, B.: Spectral graph reduction for efficient image and streaming video segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 49–56 (2014)
8. Galasso, F., Nagaraja, N.S., Cardenas, T.J., Brox, T., Schiele, B.: A unified video segmentation benchmark: annotation, metrics and analysis. In: IEEE International Conference on Computer Vision, pp. 3527–3534 (2013)
9. Garcia, D.: Robust smoothing of gridded data in one and higher dimensions with missing values. *Comput. Stat. Data Anal.* **54**(4), 1167–1178 (2010)
10. Grundmann, M., Kwatra, V., Han, M., Essa, I.A.: Efficient hierarchical graph-based video segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
11. Huang, H., Liu, H., Zhang, L.: Videoweb: space-time aware presentation of a video-clip collection. *IEEE J. Emerg. Sel. Top. Circ. Syst.* **4**(1), 142–152 (2014)
12. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: IEEE International Conference on Computer Vision, pp. 2192–2199 (2013)
13. Liu, D., Yu, Z.: A computationally efficient algorithm for large scale near-duplicate video detection. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015, Part II. LNCS, vol. 8936, pp. 481–490. Springer, Heidelberg (2015)
14. Liu, J., Huang, Z., Cai, H., Shen, H.T., Ngo, C., Wang, W.: Near-duplicate video retrieval: current research and future trends. *ACM Comput. Surv.* **45**(4), 44:1–44:23 (2013)
15. Mahmoud, K.M., Ghanem, N.M., Ismail, M.A.: Unsupervised video summarization via dynamic modeling-based hierarchical clustering. *Int. Conf. Mach. Learn. Appl.* **2**, 303–308 (2013)
16. Nock, R., Nielsen, F.: Statistical region merging. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(11), 1452–1458 (2004)
17. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(6), 1187–1200 (2014)
18. Ommer, B., Mader, T., Buhmann, J.M.: Seeing the objects behind the dots: recognition in videos from a moving camera. *Int. J. Comp. Vis.* **83**(1), 57–71 (2009)
19. Phan, R., Chia, J., Androutsos, D.: Unconstrained logo and trademark retrieval in general color image databases using color edge gradient co-occurrence histograms. *IEEE Int. Conf. Acoust. Speech, Sign. Proces.* **114**(1), 1221–1224 (2008)
20. Sadanand, S., Corso, J.J.: Action bank: a high-level representation of activity in video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1234–1241 (2012)
21. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.* **3**(1), 1–37 (2007)
22. Wang, H., Kläser, A., Schmid, C., Liu, C.: Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Computer Vision* **103**(1), 60–79 (2013)

23. Xu, C., Xiong, C., Corso, J.J.: Streaming hierarchical video segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 626–639. Springer, Heidelberg (2012)
24. Zaharieva, M., Breiteneder, C.: Recurring element detection in movies. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 222–232. Springer, Heidelberg (2012)
25. Zeppelzauer, M., Mitrovic, D., Breiteneder, C.: Analysis of historical artistic documentaries. In: International Workshop on Image Analysis for Multimedia Interactive Services, pp. 201–206 (2008)
26. Zhang, L., Xu, Q., Nie, L., Huang, H.: Videograph: A non-linear video representation for efficient exploration. *Vis. Comput.* **30**(10), 1123–1132 (2014)

MultiMedia Modeling

22nd International Conference, MMM 2016, Miami, FL,  
USA, January 4-6, 2016, Proceedings, Part I

Tian, Q.; Sebe, N.; Qi, G.-J.; Huet, B.; Hong, R.; Liu, X.  
(Eds.)

2016, XXIV, 927 p. 354 illus. in color., Softcover

ISBN: 978-3-319-27670-0