Considering Human Perception and Memory in Interactive Multimedia Retrieval Evaluations

Luca Rossetto[1 [0000−0002−5389−9465]], Werner Bailer[3 [0000−0003−2442−4900]], and Abraham Bernstein[1 [0000−0002−0128−4602]]

[1] University of Zurich, Zurich, Switzerland
{lastname}@ifi.uzh.ch
[2] JOANNEUM RESEARCH, Graz, Austria
werner.bailer@joanneum.at

Abstract. Experimental evaluations dealing with visual known-item search tasks, where real users look for previously observed and memorized scenes in a given video collection, represent a challenging methodological problem. Playing a searched "known" scene to users prior to the task start may not be sufficient in terms of scene memorization for re-identification (i.e., the search need may not necessarily be successfully "implanted"). On the other hand, enabling users to observe a known scene played in a loop may lead to unrealistic situations where users can exploit very specific details that would not remain in their memory in a common case. To address these issues, we present a proof-of-concept implementation of a new visual known-item search task presentation methodology that relies on a recently introduced deep saliency estimation method to limit the amount of revealed visual video contents. A filtering process predicts and subsequently removes information which in an unconstrained setting would likely not leave a lasting impression in the memory of a human observer. The proposed presentation setting is compliant with a realistic assumption that users perceive and memorize only a limited amount of information, and at the same time allows to play the known scene in the loop for verification purposes. The new setting also serves as a search clue equalizer, limiting the rich set of present exploitable content features in video and thus unifies the perceived information by different users. The performed evaluation demonstrates the feasibility of such a task presentation by showing that retrieval is still possible based on query videos processed by the proposed method. We postulate that such information incomplete tasks constitute the necessary next step to challenge and assess interactive multimedia retrieval systems participating at visual known-item search evaluation campaigns.

Keywords: Retrieval Evaluation, Query Generation, Interactive Retrieval, Human Perception and Memory

1 Introduction

When evaluating retrieval approaches, the different campaigns aim to simulate a realistic search scenario in a controlled environment. In the case of interactive

video retrieval, one of these scenarios is that a user of a search system has seen a specific part of a video in the past, of which they know that it is contained within a given dataset, and wants to retrieve this exact video segment as effectively as possible. This scenario can be simulated by showing such a video segment to users in a controlled environment and have them simultaneously search for it using different systems, implementing different approaches. Such a setup does however not accurately represent the conditions of the original scenario, since the users participating in the evaluation are not only aware that they will need to retrieve the video at the time they see it, but they also know the properties of the retrieval systems they are to use for that task. The evaluation participants therefore have the opportunity to pay special attention to certain aspects of the video which can be most effectively used for a query in a particular system, even though they might not have paid any attention to these aspects when looking at the same video outside of an evaluation setting. An example of how this can be exploited is shown in Figure 1, which shows the queries of two visual known-item search tasks used in the 2019 Video Browser Showdown (VBS) [34]. The red highlights emphasize legible text which was successfully used to efficiently retrieve the relevant segment during the evaluation. This text is however of minor semantic importance to the events shown in the video and would therefore probably not be remembered – or even perceived – by somebody who saw the video in an unrelated context in the past and now wants to retrieve these sequences. The task during the evaluation setting is therefore arguably not able to accurately simulate the real-world setting which is to be evaluated. To overcome this, we argue that it is insufficient to present a multimedia document directly as a query for known-item search tasks. Rather, when presenting queries to evaluation participants, one needs to consider the effects of human attention as it would likely operate in an actual unconstrained setting without specific priming which would lead to an inaccurate mental representation of the relevant document, as well as human memory effects which would, over time, alter these mental representations even further. As a proof-of-concept and first step into this direction, we propose a saliency-based filtering approach for the generation of evaluation queries, which limits the information in the video to the aspects to which a user would likely have paid attention in a regular setting. It does this by removing specific details to avoid their exploitation in a query. The method thereby implicitly predicts and subsequently removes information a human would *not* pay attention to and would therefore not remember afterwards.

After discussing some relevant related work in Section 2, we introduce the proposed proof-of-concept method and its implementation in Sections 3 and 4 respectively. Section 5 outlines the evaluation procedure and Section 6 shows its results. Finally, Section 7 concludes and offers some outlook.

2 Related Work

Video is often used as a metaphor for human memory on anything from a personal to a societal level. The relation between video and memory is a manifold

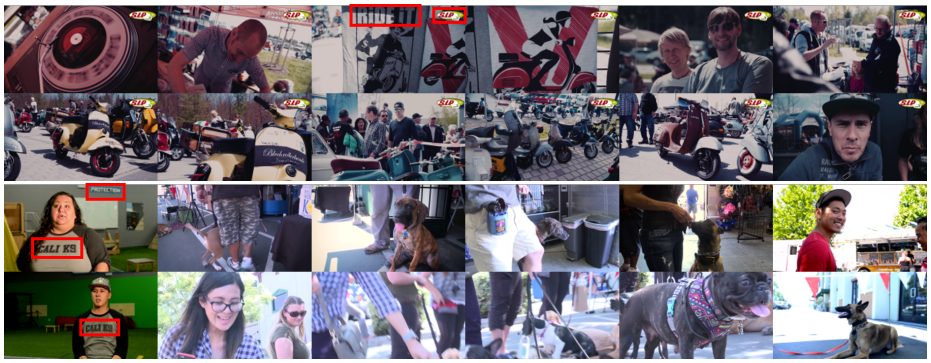


Fig. 1: Example key-frames from two visual known-item search tasks of the 2019 video browser showdown, referencing V3C [35] videos 03482 (top) and 02380 (bottom). The red boxes highlight text which was successfully used to retrieve the relevant sequence during the VBS, even though it would probably not be remembered in a real-world setting.

one, instigating research in the natural and social sciences [5] as well as the humanities [22], most of which is outside the scope for this work. In practice, however, human memory is often far less precise than video, being not only affected by forgetfulness after the fact but already impeded by selective perception and inattentional blindness [39], which can lead an observer to not even *see* certain aspects of a scene in case their attention is otherwise occupied. For query presentation, we aim to mimic the situation that the searcher remembers a scene view in video or real life some time ago. We are thus interested in long term memory of visual information, discarding cues in the query video that are not memorable. Mandler and Ritchey [28] found that humans can remember information well if can be organised in visual schema, i.e., represented in terms of their properties and arrangement. Thus anything that is coincidentally in the video, but not related to the main person/object/action of interest should thus be suppressed.

In film, an individual's memory is commonly but not exclusively depicted in *Flashbacks* [17] which are sequences outside of the temporal order of the main narrative, usually delineated with some visual transition. During a flashback or other memory sequence, various visual cues are used to help the audience identify the memory sequence. These queues usually degrade the visual fidelity of the presentation with film noise or sepia effects, vignetting or the reduction of color to monochrome [30]. Sometimes, localized effects are used to highlight particular parts of the memory, such as the selective saturation or desaturation of objects or people [16].

While human memory has a large capacity to capture multi-modal impressions in great detail [8], not all of them are equally *memorable*. Research on the *memorability* of images has shown that the likelihood of an image being remembered by a human observer is largely independent of the observers them-

selves [18] and can be estimated with a high degree of reliability. [21] presented such a memorability estimation method with a near human level rank correlation, concluding that *"predicting human cognitive abilities is within reach for the field of computer vision"*. The method shows that spatially concentrated saliency, providing a 'point of focus' increases the memorability of an image and that image memorability is positively correlated with (human) body parts and faces while being negatively correlated with natural scenes. [6] meanwhile shows that while human faces are generally a memorable part of an image, some faces are more memorable than others and that this difference is consistent across different observers. A study of the effect on overall image memorability based on different objects being visible is presented in [14]. It again confirms the positive correlation between localized saliency and memorability and shows that objects which appear towards the center of an image are more memorable, which validates the vignetting effect as an illustration for memory discussed above. The study also finds that certain 'object categories' such as people, animals or vehicles are inherently more memorable than buildings or furniture.

Less research has yet been conducted in the area of the memorability of video [38,11], which is also considerably more difficult due to its temporal aspects and multi-modal nature. Most recent research activities in this area have clustered around a recently introduced MediaEval[3] task on short- and long-term video memorability prediction [10]. Several participating teams [40,9,41,37] found that image memorability estimation do not directly translate to video memorability and that the results are worse for the long-term estimates than for the short-term ones, indicating that the temporal and multi-modal aspects of video have non-negligible effects on memory formation. All of the proposed methods also do consider the video scene as a whole and do not aim to identify or isolate the aspects which are especially memorable. This would however be a requirement for isolating especially memorable components of a video. In contrast, the concept of *saliency* can be applied locally and describes how much any particular region, in this case of an image, stands out with respect to its neighbors. Based on the observation that most existing work on memorability treats entire media items, Akgunduz et al. [3] performed an experiment in which participants were asked to identify the regions they thought helped them remembering an image. The authors found that the regions were more consistent across participants for correctly remembered images than those for false positives. They found however low overlap of these regions with an image saliency method they used for comparison. The authors thus used the data from their experiments to train a CNN for predicting regions impacting memorability.

Saliency estimation is a common task in image processing with various applications in computer vision as well as image and video coding [13]. Various methods have been proposed for the estimation of saliency in images and later videos, using both engineered and learned features. For this work, we use a recently proposed deep-learning based method [19] which is temporally consistent and has a high accuracy when compared to a human eye-gaze ground truth.

[3] `http://www.multimediaeval.org/`

In contrast to the visual domain, comparatively little work has been done in acoustic saliency or memorability detection. While there are methods for estimating both the saliency [33] and the memorability [32] of an auditory signal, they detect the salient segments of a signal (i.e., operate in the time domain) as a whole and do not isolate the salient aspects of a signal (i.e., in the frequency domain) analogously to a saliency heat map which can be produced from an image. Some novel multi-modal audio separation methods [42,15] which jointly consider aural and visual information could however form a basis for future extensions in such a direction.

3 Methodology

The proposed method aims at providing a first-order approximation of the effects of human visual attention and, by extension, memory by removing non-salient information from the video. The reasoning behind this approach is that, if an observer does not deliberately focus on specific details they know to be useful for a particular task, they will by default (meaning in the absence of such a task) focus only on the most inherently salient aspects of the video, which in turn will be the only aspects they would be able to remember at a later date. Despite saliency only being one predictive component of memorability, (as discussed in Section 2), we use it over a direct attempt at memorability estimation, since saliency can be locally estimated with a high temporal consistency, which is not currently feasible for memorability but required for filtering. We also argue that many factors influence if a salient region caused an impression sufficient to form a lasting memory, some of which might be different from individual to individual and therefore not feasibly predictable by any one model. A non-salient region however does more reliably predict a part of the video to which, little attention would be payed in an unconstrained setting. We therefore use a low saliency estimate for a region as a prediction, that this particular region would ultimately *not* be remembered, which justifies the removal of its contents.

To estimate the visual saliency, the method proposed in [19] is used, which uses a deep convolutional neural network architecture to predict a heat map of eye-gaze information, based on a sequence of consecutive video frames. Since the human visual system favors attention to the center of their visual field, the estimation of eye-gaze serves as an ideal basis for our approach.

The full processing pipeline is illustrated in Figure 2. The block labelled *Saliency Estimation* in the top row predicts an eye-gaze heat map to be used as an input of a mixer, which overlays the unmodified salient foreground over a desaturated and blurred version of the input video, used as a background. This visual degradation process is inspired by visual effects used in film to indicate flashbacks or other memory scenes and is designed in such a way as to give some visual context to the unmodified parts of the image, without providing any semantic information. We choose this degradation process over stylization methods such as [12] in order to generate a smoother transition between the filtered and

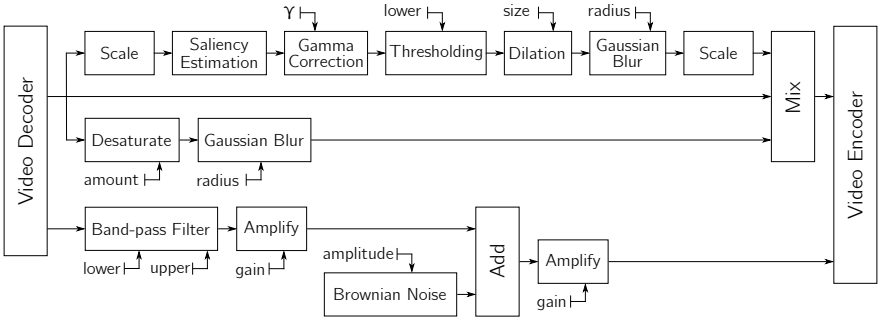


Fig. 2: Illustration of the video filtering pipeline, the upper part describing the filtering of visual information and the lower part describing the filtering of the audio signal. The used input parameters are: $\gamma = 0.8$, lower threshold = 0.5, dilation size = 4 pixel, blur radius = 5 pixel, saturation = 10%, background blur radius = 3%, lower frequency cut-off = 100Hz, upper frequency cut-off = 2kHz, noise amplitude = 0.005, amplifier gain = 20%. Parameters were determined empirically.

the unfiltered parts of the video and also to remove additional color- and shape information, since such information could still be used for query formulation.

Before the eye-gaze prediction can be used as an input mask, it is passed through a gamma correction and threshold step to extenuate regions with a high probability and remove regions with a low probability of being looked at directly. The resulting mask is dilated and blurred to ensure a smoother transition between foreground and background during mixing before being scaled to the full size of the input video.

Since isolating the salient aspects of an audio signal appears to be infeasible, we instead apply a content independent filter to the audio signal. The filter primarily consists of a band-pass filter which heavily attenuates all frequencies outside of a narrow range similar to the one used by analog telephones. The remaining pipeline is concerned with the introduction of some Brownian noise in order to hide some additional details as well as steps for amplitude adjustment. In addition, the encoder uses MP3 as output audio format and is set to the lowest supported bit-rate in order to further reduce the audio content based on the encoders internal heuristics.

Figure 3 shows examples of the filter in action. It can be seen, that areas which capture the visual attention are largely preserved while many details, like the person in the background of the first image, the signs and posters in the second image or the captions in the third image become unrecognizable.

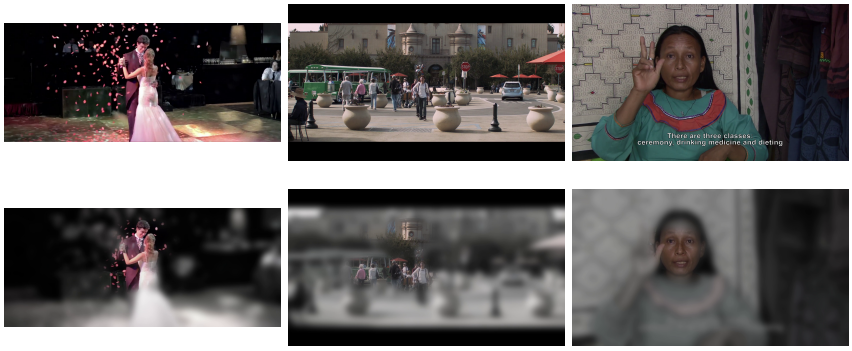


Fig. 3: Examples of the proposed method: original frames from V3C videos (left to right) 00801, 06777 and 07453 above and filtered versions of the same frames below.

4 Implementation

We implemented the filtering pipeline as a standalone application in Java, using TensorFlow [1] for the evaluation of the saliency estimation neural network, BoofCV [2] for image processing and FFmpeg[4] for video decoding and encoding. A pre-trained instance of the neural network[5] was provided by the authors of [19]. The parameters of the filter pipeline indicated in Figure 2 can be freely adjusted using an external configuration file. We provide the implementation as open source software via GitHub.[6]

5 Evaluation

We evaluated the proposed method at the 9^{th} Video Browser Showdown (VBS) co-located with the 2020 International Conference on Multimedia Modeling during a dedicated, private session using 10 different interactive video retrieval systems [4,20,23,24,25,26,27,29,31,36] in one dedicated evaluation session. This session was split into two tracks with 10 participants in each, leaving one participant per system. In both of the two tracks, participants were given 6 visual known-item search queries taken from the first shard [7] of the V3C dataset [35]. For both tracks, 3 of the queries were processed with the proposed method while the others were kept unmodified. The modifications were alternated between the two tracks, and the queries were presented in the same order in both sessions. The tasks during this session were equivalent to the regular visual known-item search tasks of VBS during which participants are given 5 minutes to find the presented video sequence of 20 seconds in a video dataset of roughly 1,000 hours.

[4] `https://ffmpeg.org/`
[5] `https://github.com/remega/OMCNN_2CLSTM`
[6] `https://github.com/lucaro/VideoSaliencyFilter`

The 6 used query videos were randomly selected from the dataset and only checked for visual diversity and uniqueness. The videos did not contain any easily identifiable or reproducible components, such as distinctive visible text or spoken dialogue.

6 Results

During the evaluation session, the 20 participants made a total of 66 submissions for the 2×6 evaluated tasks, 33 of which were correct. This number of submissions is substantially lower than what we expected, based on the data of previous VBS evaluations, independently of the application of the filtering method. This might be caused by the fact that during this special evaluation session, each participating search system was only operated by one person at a time, rather than the usual two in regular VBS settings. Of the 33 correct submissions, 15 were made for queries with unmodified videos while 18 submissions were made correctly for the queries processed by the proposed filtering approach. Figure 4 shows a breakdown of the number of correct and incorrect submissions with respect to task and filtering. There is no clearly discernible pattern relating the number of correct or incorrect submissions per task with the application of the proposed filter.

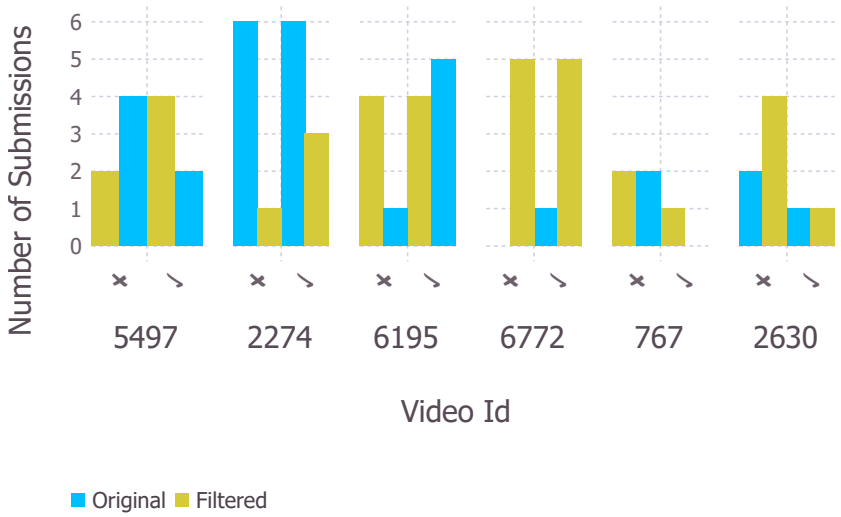
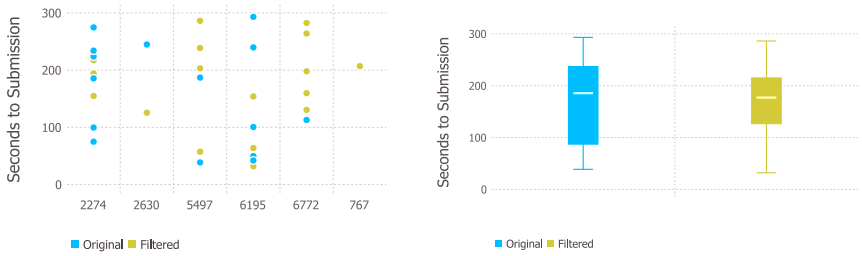


Fig. 4: Number of correct (✓) and incorrect (✗) submissions per video

Focusing only on the correct submissions, Figure 5a shows the individual submission times per task. While there are differences in submission times within

the different tasks, there appears to be no substantial overall difference with respect to the filter. The aggregated distributions of submission times of correct submissions with respect to the application of the filter are illustrated in Figure 5b. There again appears not to be any substantial difference depending on the filter.



(a) Times of correct submission per video

(b) Time distributions of correct submissions

Fig. 5: Times of correct and incorrect submissions per type

Based on the results presented above, we cannot see any substantial difference in retrieval performance between the tasks using the unfiltered videos and those using the filtered ones, which leads us to conclude that successful retrieval of the target sequence is still possible using the filtered videos and that the filtering hence does not negatively impact the solubility of the retrieval task. The fact that none of the 6 video segments happened to contain any easily exploitable components, such as recognizable text or distinctive dialogue supports the assumption that the filter leaves sufficient information intact, seeing that such aspects would have been removed by the proposed method, as illustrated in Figure 3. It is therefore reasonable to assume that the filtering could be used in an interactive video retrieval evaluation setting without negatively impacting the task as it is intended. Due to the small number of results only limited conclusions can be drawn. A larger-scale evaluation would be needed to make any strong quantitative statements about the different effects of the method on different types of query videos.

7 Conclusion & Outlook

In this paper, we presented a saliency based method for the generation of query videos containing only partial information for use in the evaluation of interactive video retrieval systems. This filtering method serves as a proof-of-concept for the feasibility of considering human perception and memory effects in the context of the evaluation of interactive multimedia retrieval approaches. The performed experiments indicate that the proposed method, while predicting and subsequently

removing many distinctive details from the video, to which presumably *little to no* attention would have been paid outside of an explicit retrieval scenario, does not negatively impact retrieval performance when compared to unfiltered videos. The results, therefore, indicate that such a method could be used in interactive video retrieval evaluation campaigns to more accurately simulate the desired real-world use case.

While the presented method serves as a first feasibility demonstration, the problem of task generation for such retrieval evaluations is however far from solved. For an accurate simulation of the scenario of a human user trying to use a retrieval system in order to find a multimedia document they encountered before and only partially remember, additional aspects such as longer-term human memory effects would need to be taken into account, which are not considered by this method. Further research in the area of multi-modal memorability estimation and, especially, localization, as well as the necessary multimedia decomposition methods is needed in order to more accurately isolate the relevant aspects of a query document in order to consistently 'implant' the information need which is supposed to serve as a basis for the relevant evaluation task into a user.

Acknowledgements

The authors would like to thank all the participants of the 2020 Video Browser Showdown who contributed to the dedicated evaluation of the queries produced using the approach presented in this paper.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). pp. 265–283 (2016), https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
2. Abeles, P.: Boofcv v0.25. http://boofcv.org/ (2016)
3. Akagunduz, E., Bors, A., Evans, K.: Defining image memorability using the visual memory schema. IEEE transactions on pattern analysis and machine intelligence (2019)
4. Andreadis, S., Moumtzidou, A., Apostolidis, K., Gkountakos, K., Galanopoulos, D., Michail, E., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: Verge in vbs 2020. In: International Conference on Multimedia Modeling. pp. 778–783. Springer (2020)
5. Bainbridge, W.A., Hall, E.H., Baker, C.I.: Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. Nature communications **10**(1), 1–13 (2019)
6. Bainbridge, W.A., Isola, P., Oliva, A.: The intrinsic memorability of face photographs. Journal of Experimental Psychology: General **142**(4), 1323 (2013)

7. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3c1 dataset: An evaluation of content characteristics. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 334–338 (2019)
8. Brady, T.F., Konkle, T., Alvarez, G.A., Oliva, A.: Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences **105**(38), 14325–14329 (2008)
9. Chaudhry, R., Kilaru, M., Shekhar, S.: Show and recall@ mediaeval 2018 vimemnet: Predicting video memorability (2018)
10. Cohendet, R., Demarty, C.H., Duong, N., Sjöberg, M., Ionescu, B., Do, T.T.: Mediaeval 2018: Predicting media memorability task. arXiv preprint arXiv:1807.01052 (2018)
11. Cohendet, R., Yadati, K., Duong, N.Q., Demarty, C.H.: Annotating, understanding, and predicting long-term video memorability. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 178–186. ACM (2018)
12. DeCarlo, D., Santella, A.: Stylization and abstraction of photographs. ACM Trans. Graph. **21**(3), 769–776 (Jul 2002). https://doi.org/10.1145/566654.566650
13. Deng, X., Xu, M., Jiang, L., Sun, X., Wang, Z.: Subjective-driven complexity control approach for hevc. IEEE Transactions on Circuits and Systems for Video Technology **26**(1), 91–106 (2015)
14. Dubey, R., Peterson, J., Khosla, A., Yang, M.H., Ghanem, B.: What makes an object memorable? In: Proceedings of the ieee international conference on computer vision. pp. 1089–1097 (2015)
15. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018)
16. Fletcher, D.: Rocketman. Paramount Pictures (may 2019)
17. Hayward, S.: Cinema Studies: The Key Concepts (Routledge Key Guides), chap. Flashback. Routledge (sep 2000)
18. Isola, P., Xiao, J., Parikh, D., Torralba, A., Oliva, A.: What makes a photograph memorable? IEEE transactions on pattern analysis and machine intelligence **36**(7), 1469–1482 (2013)
19. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: The European Conference on Computer Vision (ECCV). pp. 602–617 (September 2018)
20. Jónsson, B.Þ., Khan, O.S., Koelma, D.C., Rudinac, S., Worring, M., Zahálka, J.: Exquisitor at the video browser showdown 2020. In: International Conference on Multimedia Modeling. pp. 796–802. Springer (2020)
21. Khosla, A., Raju, A.S., Torralba, A., Oliva, A.: Understanding and predicting image memorability at a large scale. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2390–2398 (2015)
22. Kilbourn, R.: Memory and the flashback in cinema (Jul 2013). https://doi.org/10.1093/obo/9780199791286-0182
23. Kim, B., Shim, J.Y., Park, M., Ro, Y.M.: Deep learning-based video retrieval using object relationships and associated audio classes. In: International Conference on Multimedia Modeling. pp. 803–808. Springer (2020)
24. Kratochvíl, M., Veselỳ, P., Mejzlík, F., Lokoč, J.: Som-hunter: Video browsing with relevance-to-som feedback loop. In: International Conference on Multimedia Modeling. pp. 790–795. Springer (2020)

25. Le, N.K., Nguyen, D.H., Tran, M.T.: An interactive video search platform for multimodal retrieval with advanced concepts. In: International Conference on Multimedia Modeling. pp. 766–771. Springer (2020)
26. Leibetseder, A., Münzer, B., Primus, J., Kletz, S., Schoeffmann, K.: divexplore 4.0: The itec deep interactive video exploration system at vbs2020. In: International Conference on Multimedia Modeling. pp. 753–759. Springer (2020)
27. Lokoč, J., Kovalčík, G., Souček, T.: Viret at video browser showdown 2020. In: International Conference on Multimedia Modeling. pp. 784–789. Springer (2020)
28. Mandler, J.M., Ritchey, G.H.: Long-term memory for pictures. Journal of Experimental Psychology: Human Learning and Memory **3**(4), 386 (1977)
29. Nguyen, P.A., Wu, J., Ngo, C.W., Francis, D., Huet, B.: Vireo@ video browser showdown 2020. In: International Conference on Multimedia Modeling. pp. 772–777. Springer (2020)
30. Nolan, C.: Memento. Newmarket Films (sep 2000)
31. Park, S., Song, J., Park, M., Ro, Y.M.: Ivist: Interactive video search tool in vbs 2020. In: International Conference on Multimedia Modeling. pp. 809–814. Springer (2020)
32. Ramsay, D., Ananthabhotla, I., Paradiso, J.: The intrinsic memorability of everyday sounds. In: Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio. Audio Engineering Society (2019)
33. Rodriguez-Hidalgo, A., Peláez-Moreno, C., Gallardo-Antolín, A.: Echoic log-surprise: A multi-scale scheme for acoustic saliency detection. Expert Systems with Applications **114**, 255–266 (2018)
34. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., et al.: Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. IEEE Transactions on Multimedia (2020)
35. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C–A Research Video Collection. In: International Conference on Multimedia Modeling. pp. 349–360. Springer (2019)
36. Sauter, L., Parian, M.A., Gasser, R., Heller, S., Rossetto, L., Schuldt, H.: Combining boolean and multimedia retrieval in vitrivr for large-scale video search. In: International Conference on Multimedia Modeling. pp. 760–765. Springer (2020)
37. Savii, R.M., dos Santos, S.F., Almeida, J.: Gibis at mediaeval 2018: Predicting media memorability task. In: Working Notes Proceedings of the MediaEval 2018 Workshop. CEUR-WS (2018)
38. Shekhar, S., Singal, D., Singh, H., Kedia, M., Shetty, A.: Show and recall: Learning what makes videos memorable. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 2730–2739 (2017)
39. Simons, D.J., Chabris, C.F.: Gorillas in our midst: Sustained inattentional blindness for dynamic events. perception **28**(9), 1059–1074 (1999)
40. Smeaton, A.F., Corrigan, O., Dockree, P., Gurrin, C., Healy, G., Hu, F., McGuinness, K., Mohedano, E., Ward, T.E.: Dublin's participation in the predicting media memorability task at mediaeval 2018 (2018)
41. Wang, S., Wang, W., Chen, S., Jin, Q.: Ruc at mediaeval 2018: Visual and textual features exploration for predicting media memorability. In: Working Notes Proceedings of the MediaEval 2018 Workshop. CEUR-WS (2018)
42. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 570–586 (2018)