An Interactive Learning System for Large-Scale Multimedia Analytics

Omar Shahbaz Khan
IT University of Copenhagen
Copenhagen, Denmark
omsh@itu.dk

ABSTRACT

Analyzing multimedia collections in order to gain insight is a common desire amongst industry and society. Recent research has shown that while machines are getting better at analyzing multimedia data, they still lack the understanding and flexibility of humans. A central conjecture in Multimedia Analytics is that interactive learning is a key method to bridge the gap between human and machine. We investigate the requirements and design of the Exquisitor system, a very large-scale interactive learning system that aims to verify the validity of this conjecture. We describe the architecture and initial scalability results for Exquisitor, and propose research directions related to both performance and result quality.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; **Multimedia databases**.

KEYWORDS

Multimodal Retrieval; Multimedia Analytics; Interactive Learning; Exquisitor.

ACM Reference Format:
Omar Shahbaz Khan. 2020. An Interactive Learning System for Large-Scale Multimedia Analytics. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3391935

1 INTRODUCTION

Over the course of the last decade the scale of multimedia collections for industry and society has grown tremendously. This growth raises concern for people that have a desire to gain insight from these collections through data processing. In 2010, a new field has emerged called Multimedia Analytics [5], which focuses on addressing these concerns via collaboration between human and machine, complementing the strengths of each to explore and search through a collection. A central conjecture in Multimedia Analytics is that *interactive learning is a key method for performing analytical tasks on large multimedia collections* [38]. In order to verify this conjecture, however, there is a need for an interactive learning system that works at scale.

Interactive learning approaches have been around for many decades, originally used in text retrieval [14] and later in content-based image retrieval [30]. Systems based on these original methods, however, were not designed to handle the explosive scale of multimedia data of today. The state-of-the-art large-scale interactive learning approach, Blackthorn [37], is capable of interacting with collections of up to 100 million items through the use of a novel compression scheme and multicore processing achieving roughly 1.2 seconds of response time per interaction with 16 CPU cores. However, even this approach cannot be considered appropriate for investigating the conjecture as the approach has a direct correlation between scale and computational resources, meaning that if the size of the collection grows, more computational resources are required to maintain the response time. To investigate the validity of the conjecture a system must be developed that scales better and requires less computing resources.

The goal of this project is to explore algorithms and data structures required to build such a scalable interactive learning system. In particular, the project will make the following contributions:

- We have studied the effects of incorporating high-dimensional indexing into the interactive learning approach as a way to effectively reduce the search space (Section 3.1).
- We will explore how to dynamically determine the scope of retrieval, depending on quality of results or presence of filters (Section 4.1).
- We will investigate how to utilize multiple modalities to improve the result quality in the interactive learning process (Section 4.2.1).
- While investigating the above contributions, we also participate in interactive system evaluation efforts, thus partially testing the validity of the conjecture (Section 4.2.2).

The remainder of this paper is organized as follows. Section 2 outlines the background needed for the proposed work. Section 3 describes the research conducted so far during the project, while Section 4 outlines the research challenges that will be addressed to complete the project. Section 5 then concludes the paper.

2 BACKGROUND

Through recent advances in deep learning, machines are capable of extracting information from multimedia items far beyond human capacity, however, their understanding of the information is still worse than humans. This is known as the semantic gap. Furthermore the machine is restrictive and not able to adapt on the fly towards new knowledge like the human mind. This is known as the pragmatic gap. In an effort to reduce these gaps a human in the loop

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3391935

approach has been proposed in the field of Multimedia Analytics, the combination of visual analytics and multimedia analysis [5, 38]. This approach focuses on the cooperation between the machine and human to explore and search through collections to gain insight.

In [38], the tasks that a user can perform when interacting with a collection are mapped onto an exploration-search axis. A workflow for users encountering a collection could be the following: they start with exploration tasks, such as browsing, then as more insight is gained about the contents they perform search tasks, such as formulating queries, and while searching they discover another area that is worth exploring, going back to exploration tasks. Workflows such as these that alternate between exploration and search through interactions is what Multimedia Analytics is aiming to support.

There is a great emphasis on interaction in order to deal with the semantic and pragmatic gaps, however, as we are working with ever growing multimedia collections a scale gap also needs to be addressed. While there are some viable scalable solutions for solving search related tasks, there is a need for solutions that can solve scalable exploration tasks. Furthermore a solution that can deal with both type of tasks is desired [19].

With these gaps in mind, Zahálka et al. identified six requirements that must be addressed to satisfy the Multimedia Analytics workflows [37]. These requirements can be divided into two groups, **performance** and **information relevance**. The former group focus on reducing response time of interactions with the system to be within seconds (*interactivity*), while still being able to interact with very large-scale collections (*scalability*) and use modest computational resources during interactions (*availability*). The latter requirements group focus on always showing relevant items (*relevance*), using features that have a semantic meaning which the user can understand (*comprehensibility*) and avoid enforcing a data structure or hierarchy that is restrictive towards the user (*adaptability*). To investigate the conjecture for Multimedia Analytics we need a system capable of effectively addressing these requirements.

In the remainder of this section we first take a deeper look at interactive learning, an approach that focuses on learning models through human interactions. We then discuss high-dimensional indexing as a way to enhance performance and media modalities for improving information relevance.

2.1 Interactive Learning

In an interactive learning process, the system present suggestions from a model or a query that a user provides feedback on. Typically a pure interactive learning system starts by presenting the user an arbitrary sample of items from its collection and then the interactive learning process begins with subsequent interaction rounds showing results based on the model.

Interactive learning methods have been used to improve queries or classification models to access document collections [1, 14, 17] and have played an essential part in multimedia research for content-based retrieval [25, 30].

The first form of interactive learning is active learning [33], which uses interaction rounds to present the *least* confident items of the current model. After either a set amount of rounds or when the user is unable to provide feedback, the model is then used to search the collection. Active learning is a great way to train models using few items, but the approach conflicts with the *relevance* requirement as it does not show relevant items to the user during interactions.

The second form of interactive learning, which better satisfies the *relevance* requirement is user relevance feedback. This approach aims to improve a query or model through user feedback on the models current *most* confident items. This allows the user to quickly determine the quality of the model and adjust accordingly. In addition it can reduce the number of interaction rounds to gain insight.

Both approaches, active learning and user relevance feedback, have pros and cons but for our system we aim to use user relevance feedback as it satisfies the *relevance* requirement by making the user aware of the direction of the exploration with every interaction.

Blackthorn [37] is the state-of-the-art large-scale interactive learning approach. Through adaptive data compression and feature selection, multi-core processing, and a Linear SVM classification model capable of scoring items directly in the compressed domain, it is able to achieve higher precision on YFCC100M, the largest collection in the multimedia research field, over other state-of-the-art approaches that utilize nearest neighbor approaches [16]. Blackthorn addresses the information relevance requirements quite well but the fulfillment of the performance requirements are arguable. Arguably it does address the *interactivity* requirement on the YFCC100M collection [27], with a response time of 1.2 seconds per interaction, while using 16 CPU cores and 5 GB of main memory which fulfils the *availability* requirement. However, as there is a strong correlation between computational resources and stable response time in the approach it breaks the *scalability* requirement when larger collections are used, and if lesser hardware is used the *interactivity* requirement breaks. While Blackthorn is a prominent interactive learning approach there is still need for improvements to deal with the *scalability* requirement.

2.2 High-Dimensional Indexing

Indexing is typically used in order to improve performance. To enhance the performance of interactive learning with high-dimensional indexing, we have identified the following requirements [20]:

- **R1** *Short and Stable Response Time:* Achieve good result quality with response time guarantees through the use of indexing.
- **R2** *Preservation of Feature Space Similarity Structure:* The space partitioning of the indexing algorithm must preserve the similarity structure on the feature space as this is used by the interactive classifier to compute relevance.
- **R3** *k Farthest Neighbours:* To get the most confident items, the index should return k farthest neighbours (k-FN).

Scalable high-dimensional indexing methods often rely on approximation through quantization, such as scalar quantization or vector quantization, or clustering. LSH is an approach utilizing random projections that act as locality preserving hash functions [2, 7]. It stores the hashed data in B buckets in L hash tables using random hash functions. In relation to the three requirements LSH and related hashing methods fail to satisfy all [20].

Vector quantization methods typically make use of clustering approaches which determine a set of representative feature vectors to use for quantization. Other clustering approaches such as Product Quantization (PQ) [16] and its variants [3, 8, 15] cluster the high-dimensional vectors into low-dimensional subspaces that

are indexed independently. While these approaches work well they often aim to improve the distribution of data within clusters often leading to very small and large clusters breaking **R1**. They also transform the space complicating the satisfaction of **R2** [20].

The extended Cluster Pruning (eCP) algorithm [11, 12], is a clustering approach that similar to vector quantization methods satisfy **R2** and **R3**. However, instead of focusing on well distributed data it attempts to balance cluster sizes for improved performance satisfying **R1**. Essentially it performs the first step of k-means clustering and produces an hierarchical tree index with the total number of clusters for each level being determined by the desired cluster size. We consider eCP to be the most promising indexing approach for large-scale interactive learning.

2.3 Media Modalities

When dealing with multimedia data we encounter many modalities, such as visual, textual, audio, etc., which all have a variety of features that can be extracted. In order to satisfy *comprehensibility*, the representation of each modality need to be in some form of semantic feature space. Utilizing multiple modalities can greatly improve retrieval results as more information is being used to get suggestions. Several techniques have been used in order to find ways to best combine the information of modalities but challenges such as data representation, fusion and co-learning still exist [4].

One way is to use statistics to determine the importance of modalities [9], while another way is to let the user in the interactive learning approach select the priority of modalities [39]. Deep learning methods can also be applied to train functions that can be queried, which do show prominent results [36], but have not been tested on very large-scale collections. Additionally there is the dilemma of early and late fusion of modalities, which depending on the type of modalities and multimedia data that is being retrieved can favor either approach [10, 34]. Another option is to use a rank aggregation scheme rather than traditional fusing [37].

Using multiple modalities impacts the *relevance* requirement, and therefore we must look into different methods for combining these modalities in order to not negatively impact the requirement.

3 CURRENT RESEARCH RESULTS

In this section, we describe our interactive learning approach and system Exquisitor. Exquisitor extends the-state-of-the-art with a greater focus on increasing scale without neglecting the human in the interactive process. With our initial work on Exquisitor we have established it to already be the most scalable interactive learning system for multimedia in terms of both performance and quality [20]. However, improvements are still needed in order for the system to be fully capable of verifying the Multimedia Analytics conjecture; these are discussed in Section 4.

3.1 Exquisitor

Exquisitor is an extension of the user relevance feedback process. The main contribution in the extension is the tightly integrated high-dimensional index of eCP. Specifically the system uses Blackthorn's compressed data representation with a Linear SVM classification model and a modified version of eCP that is capable of scoring items in the compressed space as well.

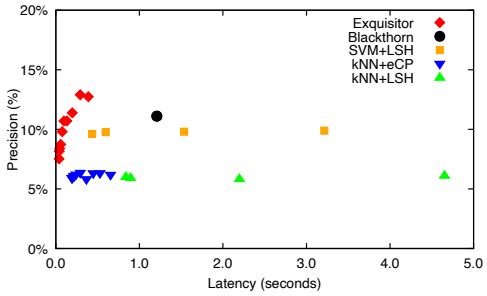

Figure 1: Average precision vs. latency over 10 rounds of analysis across all YFCC100M actors. Exquisitor, kNN+eCP: $b = 1 - 512$**. LSH:** $L = 10$, $B = [2^{10}, 2^{18}]$, $p = [15, 40]$**.**

The high-dimensional indexing of eCP adds runtime flexibility to the retrieval process. Specifically the index allows to control the scope of the search with a parameter, b, that sets the number of clusters that need to be searched through. These clusters are selected based on the trained classification model.

During retrieval, items from the selected clusters are scored for each modality using the Linear SVM followed by a rank aggregation that favors items with good scores in each modalities. Lastly k items are selected to be presented to the user.

3.2 Benchmark Evaluation

We have evaluated Exquisitor in accordance to the two requirement groups, performance and information relevance.

3.2.1 Performance. Evaluating the performance requirements for an interactive learning system is difficult as there is only one large-scale evaluation protocol available in the literature, defined over the YFCC100M collection. This protocol takes inspiration from the MediaEval Placing Task [6, 23], where actors simulate user behavior and attempt to find images from 50 different cities. The focus of the evaluation is on precision and response time. The visual features in this protocol are the 1000 ILSVRC concepts [31] extracted by GoogleNet CNN [35] and the textual features are 100 LDA topics extracted from the image metadata [28].

We have evaluated Exquisitor using this evaluation protocol comparing it with the current state-of-the-art approach Blackthorn and alternative choices for classification and indexing. The classification model of Linear SVM is compared with query based approaches using a k-NN query vector based on relevance weights [22, 29]; the high-dimensional index of eCP is compared with an multiprobe-LSH index [26]. The results of the experiments for this evaluation protocol can be seen in Figure 1. For Exquisitor (and kNN+eCP), each dot corresponds to a different b value, ranging from 1 to 512; for the LSH-based approaches, each dot is a different configuration. Figure 1 shows that Exquisitor is significantly better in both precision and response time compared to the other methods. Note that Blackthorn used 16 CPU cores, while Exquisitor requires only 1.

3.2.2 Information Relevance. The ability of a classification model to classify new semantic features not found in the current representation is vital to its information relevance. This sort of learning problem is often referred to as zero-shot learning, where a model

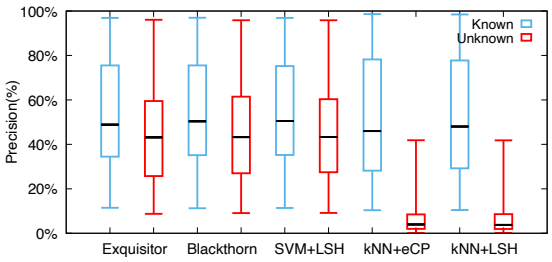

Figure 2: Results from ImageNet evaluation protocol. Blue columns depict the case where the feature is known. Red columns depict the case where the feature is unknown [20].

is trained to find a certain feature without having the knowledge of the feature during training. As there is no evaluation protocol addressing this type of problem for large-scale interactive learning system, we have designed our own protocol using the popular ImageNet dataset that contains ~14 million images. This protocol is used to compare the *adaptability* of classification models towards new knowledge. Figure 2 shows the results of this protocol. The "known" case is the baseline, where all 1000 concepts of the ILSVRC visual features are retained, while in the "unknown" case concepts are methodically removed and their images then sought without the concept. Figure 2 clearly indicates that Linear SVM performs much better than a k-NN-based approach.

3.3 Evalutation via Live Events

Designing and implementing evaluation protocols is necessary to automate the evaluation of interactive learning systems, but these protocols only allow for specific functionality to be tested. Therefore in order to evaluate our system not just in terms of performance and quality over evaluation protocols, we have to observe how humans interact with it. A way to do this is to participate in interactive system challenges, such as the Lifelog Search Challenge (LSC) [13] and Video Browser Showdown (VBS) [24, 32].

Exquisitor has so far been part of LSC 2019 [21] and VBS 2020 [18], placing 6th and 5th respectively, and was also demonstrated at ACM Multimedia 2019 [27]. Through these events we have observed that presenting arbitrary items at the start of the interactive learning process often leads to lack of positive examples, which in turn leads to additional interaction rounds. To alleviate this issue we need a way to find initial positive examples to start the learning process, such as a search function. Additionally narrowing the search space using filters on metadata, concepts, colors, etc. is useful and should be integrated. Another observation is that equal ranking of modalities is not appropriate in cases where one modality is clearly stronger than the rest. Specifically for videos, text features describe the entire video, whereas the visual features describe the shots within the video, making the textual modality add noise to the retrieval process. These are some observations made during these events that are issues which need to be addressed.

4 PROJECTED RESEARCH

Exquisitor is currently the-state-of-the-art large-scale interactive learning system. However, as we have observed through its usage in live challenges and demonstrations there are still improvements that need to be made before it is a system that can verify the conjecture. In this section, I describe the work we plan to do over the remaining course of the project.

4.1 Performance

In terms of performance, we still encounter scenarios where the emphasis on *interactivity* can negatively impact *relevance*. This happens when a user narrows the scope using strong filters that result in few or zero suggestions requiring additional rounds or restarting the process. With eCP we are capable of reducing or expanding the scope to b clusters at runtime. It is in our best interest to automatically adjust this parameter when realizing that the selected clusters will not contain enough suggestions. Furthermore, adding filtering possibilities at cluster selection levels when a more search focused task is being run can benefit performance by avoiding processing clusters that will not affect the result. We intend to implement these features in order to avoid unnecessary interaction rounds.

4.2 Information Relevance

4.2.1 Combining Multiple Modalities. Currently Exquisitor uses two modalities, visual semantic features extracted from deep neural networks and textual semantic features typically extracted as LDA topics. These work well in the interactive learning setting as the human user can relate to the features. However, additional modalities and metadata exist for multimedia data, such as time, location data, color histograms, audio features, SIFT, etc., which may also be beneficial for the learning process. Especially when dealing with multimedia data such as videos and there temporal nature, having more modalities may help to easier define the context to explore. We need to investigate the effects of adding these other modalities without negatively impacting *comprehensibility* and *relevance*.

4.2.2 Evaluating in practice. As mentioned in Section 3.3, interactive live events are important to evaluate our system. Therefore we aim to continue participating in challenges like LSC and VBS in order to get continuous feedback and better understand the interactive learning approach. These events can possibly also provide reasonable examples of user behavior that can be turned into an evaluation benchmark for interactive learning systems.

5 CONCLUSION

In this paper, we presented a central conjecture in Multimedia Analytics and the need for a truly large-scale interactive learning system in order to verify it. To that end we described Exquisitor, our interactive learning system for very large multimedia collections that we have been working on in the first half of the project. Through experiments conducted on YFCC100M and ImageNet collections we have established it to be the state-of-the-art interactive large-scale system in terms of performance, quality, and hardware requirements. Furthermore, we have observed the system in interactive search challenges and verified that the approach behind the system is valid for verifying parts of the conjecture. However, we have also established that there is still considerable work to be done, which will be done during the remainder of this research project.

REFERENCES

[1] James Allan. 1996. Incremental relevance feedback for information filtering. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. 270–278.

[2] Alexandr Andoni and Piotr Indyk. 2006. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. In *FOCS*. IEEE Computer Society, Berkeley, CA, USA, 459–468.

[3] Artem Babenko and Victor Lempitsky. 2014. The inverted multi-index. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2014), 1247–1260.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[5] Nancy A Chinchor, James J Thomas, Pak Chung Wong, Michael G Christel, and William Ribarsky. 2010. Multimedia Analysis + Visual Analytics = Multimedia Analytics. *IEEE computer graphics and applications* 30, 5 (2010), 52–60.

[6] Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. 2015. The placing task at MediaEval 2015. In *Proceedings of the MediaEval 2015 Workshop*. CEUR, Wurzen, Germany, 2.

[7] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proc. ACM Symposium on Computational Geometry*. ACM, Brooklyn, NY, USA, 253–262.

[8] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. 2013. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence* 36, 4 (2013), 744–755.

[9] King-Shy Goh, Edward Y Chang, and Wei-Cheng Lai. 2004. Multimodal concept-dependent active learning for image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*. 564–571.

[10] Iva Gornishka, Stevan Rudinac, and Marcel Worring. 2020. Interactive Search and Exploration in Discussion Forums Using Multimodal Embeddings. In *International Conference on Multimedia Modeling*. Springer, 388–399.

[11] Gylfi Þór Gudmundsson, Laurent Amsaleg, and Björn Þór Jónsson. 2012. Impact of Storage Technology on the Efficiency of Cluster-based High-dimensional Index Creation. In *Proc. International Conference on Database Systems for Advanced Applications (DASFAA)*. Springer, Busan, South Korea, 53–64.

[12] Gylfi Þór Gudmundsson, Björn Þór Jónsson, and Laurent Amsaleg. 2010. A Large-scale Performance Study of Cluster-based High-dimensional Indexing. In *Proc. International Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval (VLS-MCMR)*. ACM, Firenze, Italy, 31–36.

[13] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Dang Nguyen, Duc Tien, Michael Riegler, Luca Piras, et al. 2019. Comparing approaches to interactive lifelog search at the lifelog search challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59.

[14] Donna Harman. 1992. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. 1–10.

[15] Jae-Pil Heo, Zhe Lin, and Sung-Eui Yoon. 2014. Distance Encoded Product Quantization. In *CVPR*. IEEE Computer Society, Columbus, OH, USA, 2139–2146.

[16] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.

[17] Thorsten Joachims. 1996. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Technical Report. Carnegie-mellon univ pittsburgh pa dept of computer science.

[18] Björn Þór Jónsson, Omar Shahbaz Khan, Dennis C Koelma, Stevan Rudinac, Marcel Worring, and Jan Zahálka. 2020. Exquisitor at the Video Browser Showdown 2020. In *International Conference on Multimedia Modeling*. Springer, 796–802.

[19] Björn Þór Jónsson, Marcel Worring, Jan Zahálka, Stevan Rudinac, and Laurent Amsaleg. 2016. Ten research questions for scalable multimedia analytics. In *International Conference on Multimedia Modeling*. Springer, 290–302.

[20] Omar Shahbaz Khan, Björn Þór Jónsson, Stevan Rudinac, Jan Zahálka, Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Gylfi Þór Guðmundsson, Laurent Amsaleg, and Marcel Worring. 2020. Interactive Learning for Multimedia at Large. In *European Conference on Information Retrieval*. Springer.

[21] Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2019. Exquisitor at the lifelog search challenge 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*. 7–11.

[22] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. 2008. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence* 31, 4 (2008), 721–735.

[23] Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman, and Gareth J. F. Jones. 2011. Automatic Tagging and Geotagging in Video Collections and Communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (Trento, Italy). ACM, New York, NY, USA, Article 51, 8 pages.

[24] J. Lokoč, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad. 2018. On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015–2017. *IEEE Transactions on Multimedia* 20, 12 (2018), 3361–3376.

[25] Ye Lu, Chunhui Hu, Xingquan Zhu, HongJiang Zhang, and Qiang Yang. 2000. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the eighth ACM international conference on Multimedia*. 31–37.

[26] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. 2007. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In *Proceedings of the 33rd international conference on Very large data bases*. 950–961.

[27] Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Omar Shahbaz Khan, Björn Þór Jónsson, Gylfi Þór Guðmundsson, Jan Zahálka, Stevan Rudinac, Laurent Amsaleg, and Marcel Worring. 2019. Exquisitor: breaking the interaction barrier for exploration of 100 million images. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1029–1031.

[28] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[29] Stephen E Robertson and Karen Spärck Jones. 1994. *Simple, proven approaches to text retrieval*. Technical Report. University of Cambridge, Computer Laboratory.

[30] Yong Rui, Thomas S Huang, and Sharad Mehrotra. 1997. Content-based image retrieval with relevance feedback in MARS. In *Proceedings of International Conference on Image Processing*, Vol. 2. IEEE, 815–818.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (01 Dec 2015), 211–252.

[32] Klaus Schoeffmann. 2014. A user-centric media retrieval competition: The video browser showdown 2012-2014. *IEEE MultiMedia* 21, 4 (2014), 8–13.

[33] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[34] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 399–402.

[35] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Proc. IEEE CVPR*. IEEE Computer Society, Boston, MA, USA, 1–9.

[36] Pengcheng Wu, Steven CH Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*. 153–162.

[37] Jan Zahálka, Stevan Rudinac, Björn Þór Jónsson, Dennis C Koelma, and Marcel Worring. 2018. Blackthorn: Large-Scale Interactive Multimodal Learning. *IEEE Transactions on Multimedia* 20, 3 (2018), 687–698.

[38] Jan Zahálka and Marcel Worring. 2014. Towards interactive, intelligent, and integrated multimedia analytics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 3–12.

[39] David Zellhöfer. 2012. A permeable expert search strategy approach to multimodal retrieval. In *Proceedings of the 4th Information Interaction in Context Symposium*. 62–71.