Web-scale Multimedia Search for Internet Video Content

Lu Jiang
≪Supervised by Alexander Hauptmann and Teruko Mitamura≫
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lujiang@cs.cmu.edu

ABSTRACT

The World Wide Web has been witnessing an explosion of video content. Video data are becoming one of the most valuable sources to assess insights and information. However, existing video search methods are still based on text matching (*text-to-text* search), and could fail for the huge volumes of videos that have little relevant metadata or no metadata at all. In this paper, we propose an accurate, efficient and scalable semantic search method for Internet videos that allows for intelligent and flexible search schemes over the video content (*text-to-video* search and *text&video-to-video* search). To achieve this ambitious goal, we propose several novel methods to improve accuracy and efficiency. The extensive experiments demonstrate that the proposed methods are able to surpass state-of-the-art accuracy and efficiency on multiple datasets. Based on the proposed methods, we implement E-Lamp Lite, the first of its kind large-scale semantic search engine for Internet videos. According to National Institute of Standards and Technology (NIST), it achieved the best accuracy in the TRECVID Multimedia Event Detection (MED) 2013, 2014 and 2015, one of the most representative task for content-based video search. To the best of our knowledge, E-Lamp Lite is the first content-based semantic search system that is capable of indexing and searching a collection of 100 million videos.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Search and Retrieval**]: Search process; I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms, Experimentation, Performance

Keywords

Big Data; Web Search; Video Content Analysis; Content-based Retrieval; Multimedia Event Detection

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2888599.

1. INTRODUCTION

We are living in an era of big data: three hundred hours of video are uploaded to YouTube every minute; social media users are posting 12 millions videos on Twitter every day. According to a Cisco study, video content will account for 80% of the entire world's internet traffic by 2019. The big video data on the web are important not because there is a lot of it but because it is becoming a valuable source for insights and information, e.g. telling us about things happening in the world, giving clues about a person's preferences, pointing out places, people or events of interest, providing evidence about activities that have taken place [27].

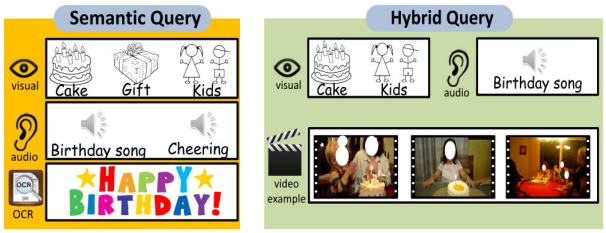
An important approach of acquiring information and knowledge is through video search. However, existing large-scale video search methods are still based on *text-to-text* matching, in which the query words are matched against the textual metadata generated by the uploader [5]. The text-to-text search method, though simple, is of minimum functionality because it provides no understanding about the video content. As a result, the method proves to be futile in many scenarios, in which the metadata are either missing or less relevant to the visual video content. According to a recent study [30], 66% videos on a social media site called Twitter Vine are not associated with meaningful metadata (hashtags or mentions), which suggests on an average day, around 8 million videos may never be watched again just because there is no way to find them. The phenomenon is more severe for the even larger amount of videos that are captured by mobile phones, surveillance cameras and wearable devices that end up not having any metadata at all. Comparable to the days in the late 1990s, when people usually got lost in the rising sea of web pages, now they are overwhelmed by the vast amounts of videos, but lack powerful tools to discover, not to mention to analyze, meaningful information in the video content.

To this end, we approach an ambitious problem called Content-Based Video Semantic Retrieval (CBVSR), in which the goal is to retrieve relevant videos not based on textual metadata, but on video content understanding. CBVSR is a type of content-based video retrieval focusing on the semantic understanding about video content. A distinguishing characteristic of CBVSR is the capability to search and analyze videos based on semantic features that are automatically extracted from the video content. Semantic features are the human interpretable multimodal features about video content such as people, objects, scenes, actions and activities, speech, visible text, etc. The CBVSR method advances traditional video retrieval methods in many ways. It enables

a more intelligent and flexible search paradigm that traditional text-to-text search would never achieve. In this paper, we consider two types of queries: a query only consisting of semantic features (e.g. people, objects, speech, visible text, etc.) is called a *semantic query*. A query consisting of both semantic features and a few video examples is called a *hybrid query*. The semantic query provides an approach for text-to-video search, and the hybrid query offers a mean for text&video-to-video search. Example 1 illustrates some examples of the queries.

EXAMPLE 1. *Suppose our goal is to search the videos about birthday party. In traditional text queries, we have to search the keywords in the user-generated metadata (titles or descriptions). For videos without any metadata, there is no way to find them. In contrast, in a semantic query, we might look for visual clues in the video content such as "cake", "gift" and "kids", audio clues like "birthday song" and "cheering sound", or visible text like "happy birthday". See Fig. 1(a). Semantic queries are flexible and can be refined by Boolean operators. For example, to capture only the outdoor party, we may add "AND outdoor' to the current query. Temporal relation between concepts can also be specified by a temporal operator. For example, we may add a temporal operator between "gift" and "cake" to find videos in which the opening of presents are seen before consuming the birthday cake.*

After watching the retrieved videos for a semantic query, the user is likely to select a few interesting videos, and to find more relevant videos like these. This can be achieved by issuing a hybrid query which adds the selected videos to the query. See Fig. 1(b). Users may also change the semantic features in the hybrid query to refine or emphasize certain aspects in the video examples. For example, we may add "AND birthday song" in the hybrid query to find more videos not only similar to the examples but also have happy birthday songs in their content.



(a) Semantic Query (b) Hybrid Query

Figure 1: Comparison of the semantic and the hybrid query on "birthday party".

The idea of CBVSR sounds appealing but, in fact, it is a very challenging problem. It introduces several novel issues that have not been sufficiently studied in the literature. As far as this study is concerned, it presents the following challenges: 1) Accurate retrieval for complex queries. A crucial challenge for any retrieval system is achieving a reasonable accuracy, especially for the top-ranked documents or videos. Unlike many problems, the data in this problem are real-world noisy and complex Internet videos, and the queries are of complex structures containing both texts and video examples. How to design intelligent algorithms to obtain state-of-the-art accuracy is a challenging issue. 2) Efficient retrieval at very large scale. Processing video proves to be a computationally expensive operation. The huge volumes of Internet video data become a key challenge for retrieval. How to design efficient algorithms that are able to search hundreds of millions of video within the maximum recommended waiting time for a user, i.e. 2 seconds [20], while maintaining maximum accuracy becomes a critical challenge.

In this paper, we propose a collection of novel methods to solve these challenges. First, we introduce a novel self-paced curriculum learning theory that allows for training more accurate semantic concepts. Second, we propose a novel and scalable approach to index semantic concepts that can significantly improve the search efficiency and scalability. Third, we design a novel video reranking algorithm that can boost accuracy for video retrieval. The proposed methods are extensively verified on a number of large-scale challenging datasets. Experimental results demonstrate that the proposed method can exceed state-of-the-art accuracy and efficiency. Furthermore, it can efficiently scale up the search to hundreds of millions of Internet videos. It only takes about 0.2 second to search a semantic query on a collection of 100 million videos, and 1 second to handle a hybrid query over 1 million videos.

The proposed methods are fundamental and can potentially benefit a number of related tasks on video search and analysis, such as video hyperlinking [22, 1], video question answering, video summarization and recommendation [4, 3], social video stream analysis [24, 16], in-video advertising [14], etc. The insight in our web-scale method may guide the design of future search or analysis systems for web-scale video data. To summarize, our contributions are as follows:

1. The first-of-its-kind framework for text-to-video and text&video-to-video semantic search over hundreds of millions of videos.
2. A novel theory about self-paced curriculum learning and its application on concept detector training.
3. A novel and cost-effective reranking algorithm.
4. A concept adjustment method that allows for efficient indexing big video data by the modified inverted index.

2. STATE OF THE ART

This section reviews some important related work on content-based video retrieval. *Content-based Image Retrieval* is a task of finding identical or visually similar images in a large collection. It provides a scheme of image-to-image search, where the query is usually a single example image. The similarity matching is based on low-level descriptors that carry little semantic meaning. Therefore it only finds visually similar, but not necessarily semantically similar images. This method can be extended to search key frames in a video clip, i.e. image-to-video search. For example, Sivic et al. introduced Video Google [26], a system to retrieve similar video key frames for a query image. Another example is searching key frames of a specific instance about, e.g., a person, a logo or a landmark [35]. Content-based image retrieval is a well-studied problem. State-of-the-art systems can efficiently handle more than 100 million images.

Semantic Concept Detection is a task of searching the occurrence of a single concept in the video content. A concept is a visual or acoustic semantic tag on objects, scenes, actions, etc. This line of study first emerged in a TRECVID task called Semantic Indexing [21]. Its subproblems like *Action Detection* and *Object Detection* [25] recently become popular very quickly. Semantic Concept Detection provides some semantic understanding about the video content but

only supports a simple search scheme like object-to-video or action-to-video search.

Multimedia Event Detection (MED): with the advance in semantic concept detection, people started to focus on searching more complex queries called events. An event is more complex than a concept as it usually involves people engaged in process-driven actions with other people and/or objects at a specific place and time [7]. For example, the event "rock climbing" involves a climber, mountain scenes, and the action climbing. A benchmark task on this topic is called TRECVID Multimedia Event Detection (MED) [23, 32]. Its goal is to provide a video-to-video search scheme. MED is a challenging problem, and the biggest collection in TRECVID only contains 200 thousand videos.

The CBVSR problem is similar to MED but advances it in the following ways. First, the queries are complex queries consisting of both text description of semantic features and video examples. Second, the search is solely based on content understanding rather than low-level features matching. Finally, the problem scale is orders-of-magnitude larger than that of MED.

3. METHOD OVERVIEW

In this paper, we model a CBVSR problem as a retrieval problem, in which given a query, we are interested in finding a ranked list of relevant videos based on the understanding about video content. To address the problem, we incorporate a two-stage framework, as illustrated in Fig. 2. The offline stage is called semantic indexing, which aims at extracting semantic features in the video content and indexing them for efficient online search. It usually involves the following steps: a video clip is first represented by the *low-level features* that capture the local appearance, texture or acoustic statistics of a video clip, represented by a collection of local descriptors [12]. State-of-the-art low-level features included in our paper are dense trajectories, Convolution Neural Network (CNN) features (GoogleNet) [28] for visual modality, and neural network features for audio modality [19]. The low-level features are then input into the off-the-shelf detectors to extract the *semantic features*. The semantic features are human interpretable features, each dimension of which corresponds to a confidence score of detecting a concept or a word in the video [7, 31]. The visual/audio concepts, Automatic Speech Recognition (ASR) [17, 18] and Optical Character Recognition (OCR) are four types of semantic features considered in this paper. After extraction, the semantic features will be adjusted and indexed for the efficient online search.

The second stage is an online stage called video search. We employ two modules to process the semantic query and the hybrid query. Both modules consist of a query generation and a multimodal search step. A user may express a query in a variety of forms such as a text description or a few video examples. The query generation for semantic query is to map the out-of-vocabulary concepts in the user query to its most relevant alternatives in the system vocabulary. For the hybrid query, the query generation also involves training a classification model using the selected video examples. The search component aims at retrieving a ranked list using the multimodal features. This step is a retrieval process for the semantic query and a classification process for the hybrid query. Afterwards, we can refine the results by reranking the videos in the returned ranked list. This process is known as reranking or Pseudo-Relevance Feedback (PRF). The basic idea is to first selects a few videos and assign assumed labels to them. The samples with pseudo labels are then used to build a reranking model using sematic and low-level features to improve the ranked list.

4. RESEARCH STUDIES

To address the challenges in the introduction, we conducted a number of studies to explore the research direction on accurate, efficient and scalable video search. Thorough the discussions of these directions, detailed methods and experimental results are provided in [13, 10, 14, 9, 8, 33, 11].

4.1 Concept Detector Construction

In [10, 9], we proposed a theory named Self-Paced Curriculum Learning (SPCL) to train robust concept detectors. The theory is inspired by the cognitive processes of humans and animals, which generally start with learning easier aspects of a task, and then gradually consider more complex examples [2, 15]. SPCL is formulated as a concise optimization problem that takes into account both prior knowledge known beforehand and the learning feedback during training. Consider a binary classification problem. Let $L(y_i, g(\mathbf{x}_i, \mathbf{w}))$ denote the loss function which calculates the cost between the ground truth label y_i and the estimated label $g(\mathbf{x}_i, \mathbf{w})$. Here $\mathbf{w}$ represents the model parameter inside the decision function g. Given a predetermined curriculum, we have:

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi) = \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i,\mathbf{w})) + f(\mathbf{v};\lambda) \quad \text{subject to } \mathbf{v} \in \Psi \tag{1}$$

where $\mathbf{v} = [v_1, v_2, \cdots, v_n]^T$ denote the latent weight variables reflecting the samples' importance. f is the self-paced function; Ψ is a feasible region that encodes the information of a predetermined curriculum. A curriculum can be expressed as a ranking function that assigns learning priorities to training samples. A self-paced function f determines a learning scheme for the model to learn new samples. Since humans use different learning schemes for different tasks, SPCL can utilize multiple learning schemes for different problems. We incorporate five types of self-paced function named binary [15, 34], linear [10], logarithmic [10], mixture [10], and diverse learning scheme [9] for concept training. The generic concept training is solved by an alternative convex search in Algorithm 1.

Experimental results on a number of benchmarks verify the proposed method can train robust detectors than the baseline methods [9, 10]. Using the methods, we have built more than 3000+ concept detectors over 2 million video clips, and incorporated them into our CBVSR system [13]. To the best our knowledge, it is by far the largest concept collection directly trained on videos. We share the semantic features at `http://www.cs.cmu.edu/~lujiang/0Ex/icmr15.html`.

4.2 Concept Adjustment and Indexing

In [14], we studied efficient concept indexing for web-scale search. We found that raw concept detection scores are inappropriate for indexing due to two types of inconsistencies. The *distributional inconsistency* means that the distribution of the raw detection score is inconsistent with the underlying concept distribution of the video. The underlying concept representation tends to be sparse but the distribution of the

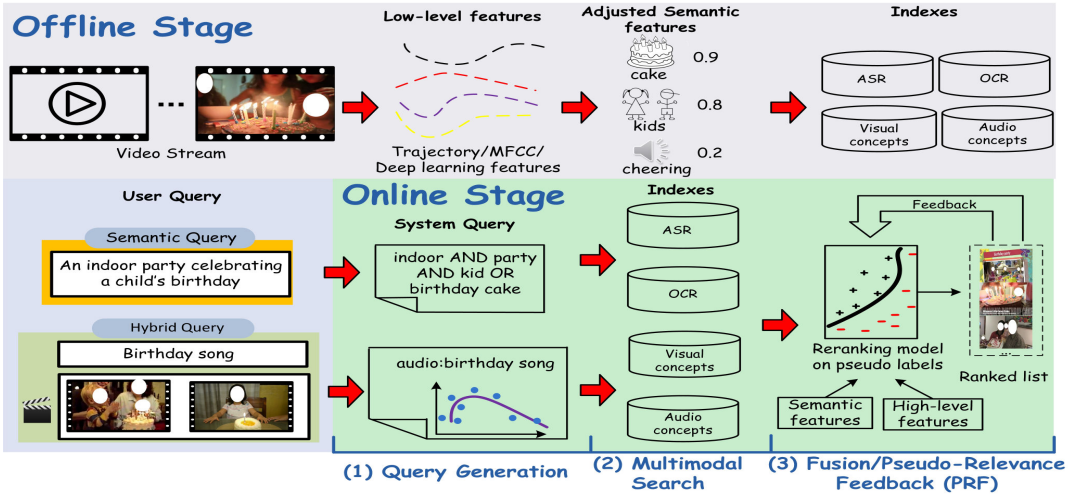


Figure 2: Overview of the framework of E-Lamp Lite.

Algorithm 1: Self-paced Curriculum Learning.

input : Input dataset $\mathcal{D}$, predetermined curriculum γ, self-paced function f and a stepsize μ
output: Model parameter $\mathbf{w}$

```
Derive the curriculum region Ψ from γ;
Initialize v*, λ in the curriculum region;
while not converged do
    Update w* = arg min_w E(w, v*; λ, Ψ);
    Update v* = arg min_v E(w*, v; λ, Ψ);
    if λ is small then increase λ by the stepsize μ;
end
return w*
```

detection score is dense, i.e. a video contains every concept. The *logical inconsistency* means that the detection scores are not consistent with the semantic relation between concepts, e.g. a video contains a "terrier" but not a "dog". The inconsistent representation can lead to inaccurate search results if not properly handled. To address the inconsistencies, we proposed a novel method called concept adjustment [14]. It aims at generating consistent concept representations that can be efficiently indexed and searched. We proposed an adjustment method to model the concept distribution and relation, in the form of an optimization problem with solid probabilistic interpretations. After adjustment, a video is represented by a few salient concepts that are logically consistent with the complex relations between concepts. We then modified the inverted inverted index structure so that it can index the adjusted concept scores. The experimental results show that the concept adjustment and indexing method provides a foundation for web-scale video search. For semantic queries, it is able to scale up the search to 100 million videos while maintaining state-of-the-art accuracy.

4.3 Query Generation & Video Search

In [13], we studied the query generation and the multimodal search for semantic queries. To map the out-of-vocabulary words in a user query to their most relevant concepts in the system vocabulary, we investigated the following mapping algorithms: exact word matching, WordNet mapping, Point-wise Mutual Information (PMI) mapping in Wikipedia, and word embedding mapping. As there is no single retrieval model that can always work the best for all semantic features, we empirically study the classical retrieval models on various types of semantic features, which includes Vector Space Model, Okapi BM25, Language Model-JM Smoothing and Language Model-Dirichlet Smoothing. Experimental results showed that the query generation and the retrieval algorithms have substantial impact on the retrieval performance. The fusion of different mapping algorithms combined with the feature-specific retrieval method yields the best results.

In [33], we studied the query generation and the search for a special type of hybrid query that only contains a few video examples. We explored a fast prediction model using Product Quantization (PQ) [6]. We proposed an approximate search that leads insignificant change in accuracy but significantly improves search efficiency. Experimental results show that the method can search 1 million videos with 1 core in less than 1 second while retaining 80% of accuracy of a state-of-the-art system [33].

4.4 Video Reranking

In [11, 8], we studied a cost-effective PRF (aka. reranking) method for both semantic and hybrid queries. We incorporate a multimodal PRF method called SPaR, which models the reranking process as a self-paced learning process [15] where the easy samples are the videos ranked at the top as they are, generally, more relevant than those videos ranked lower. To run PRF, we first need to pick a reranking model (e.g. SVM or regression model), a self-paced function [8], and reasonable starting values for the pseudo labels. The starting values can be initialized either by top ranked videos in the returned ranked lists or by other PRF methods. After the initialization, we iterate the following three steps, similar to the steps in Algorithm 1: 1) training a model based on the selected pseudo samples and their weights; 2) calculating pseudo positive samples and their weights by the self-paced function f, and selecting pseudo negative samples randomly; 3) increasing the model age to include more positive samples in the next iteration. Average fusion of the PRF result with and original ranked list is used to obtain better results. Experimental results demonstrate that SPaR can improve the performance for both hybrid and semantic queries [13, 33].

5. EXPERIMENTS

The experiments are conducted on two TRECVID benchmarks called Multimedia Event Detection (MED): MED13Test and MED14Test, the most representative benchmarks on

our problem. Each set includes 20 events over 25,000 test videos such as "changing a vehicle tire", "townhall meeting", etc. The performance is evaluated by Mean Average Precision (MAP), the official metric used by NIST. All experiments are conducted without using any text metadata. Due to the lack of space, only important results are presented.

To evaluate the accuracy, we conduct experiments using the semantic queries automatically generated using all features (Auto), using only visual features (AutoVisual), generated by human experts (Expert). We also include hybrid queries containing 10 video examples. Table 1 show the results and compare the accuracy with and without the proposed reranking method SPaR discussed in Section 4.4. It worth mentioning that the results in Table 1 are comparable or even better than the state-of-the-art accuracy on the benchmarks [13, 33]. The results verify the state-of-the-art accuracy of the proposed method. The MAP reported in Table 1 is low because the the positive to negative ratio in the benchmark is about 0.08%. NIST created the benchmarks to simulate a real-world search scenario. Even though in this challenging setting, the mean inverse rank of the first relevant video (MRR) is about 0.6. Besides, if we apply the method on another big dataset, an significant improvement over the top ranked videos can be spotted. More results can be found in [14]. We hypothesize the promising results may catalyze the rise of next generation of content-based video search, analysis and understanding.

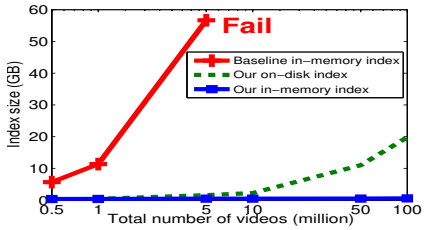
Table 1: Overview of Search Accuracy.

Query	MAP	
	MED13	MED14
Semantic Query AutoVisual	0.074	0.086
Semantic Query Auto	0.118	0.100
Semantic Query Expert	0.183	0.172
Semantic Query Expert + SPaR	0.208	0.196
Hybrid Query (10 examples)	0.258	0.220
Hybrid Query + SPaR	0.280	0.233

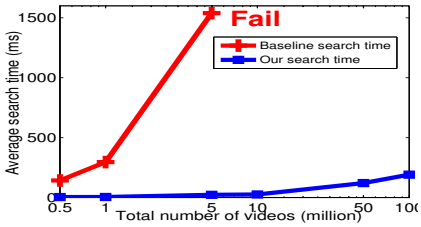
To evaluate the efficiency and scalability for semantic queries, we duplicate the videos and video shots in the largest public multimedia collection called YFCC100M [29], and create an artificial set of 100 million videos. We compare the search performance of the proposed method to a common approach in existing studies that indexes the video by dense matrices. The experiments are conducted on a single core of Intel Xeon 2.53GHz CPU with 64GB memory. The performance is evaluated in terms of the memory consumption and the online search efficiency. Fig. 3(a) compares the in-memory index as the data size grows, where the x-axis denotes the number of videos in the log scale, and the y-axis measures the index in GB. As we see, our method is scalable and only needs 550MB memory to search 100 million videos. Fig. 3(b) compares the online search speed. A similar pattern can be observed in Fig. 3 that our method is much more efficient than the baseline method and only costs 191ms to process a query on a single core. The above results verify the scalability and efficiency of the proposed method.

6. DISCUSSIONS AND FUTURE WORK

In this paper, we studied a fundamental research problem of searching semantic information in video content at a very large scale. We proposed several novel methods focusing on improving accuracy, efficiency and scalability in the novel search paradigm. The proposed methods demonstrated promising results on web-scale semantic search for video. The extensive experiments demonstrated that the methods are able to surpass state-of-the-art accuracy on multiple datasets and achieve promising results on web-scale semantic search for video.



(a) Index (in GB)



(b) Search Time (in ms)

Figure 3: The scalability test on 100 million videos.

There are several research issues to be addressed in our study. First, semantic and hybrid queries are handled by two different methods; the method for semantic queries is scalable but the one for hybrid queries is not. We extrapolate there exists a fundamental method that can unify the two methods and provides a scalable solution for hybrid queries. Second, preliminary studies do not focus on how to interpret the search results. This problem may be addressed by utilizing existing methods studied in other communities. Finally, as the proposed method provides a fundamental functionality of assessing semantic information in video, we would like to discuss how the method can benefit problems in the areas of web search and data analysis.

Back to the days in the late 1990s when people often got lost in the rising sea of web pages, the search engines, such as Google and Yahoo, were only designed to find a URL for a specific web page. However, though 15 years of evolution, search engines are becoming a foundation for various large-scale applications such as question answering, hypertext linking, and web/mobile ads. We believe video search and analysis may follow a similar path. From this perspective, the proposed method is a merely concrete step towards a more intelligent and promising future.

Acknowledgments

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number OCI-1053575. It used the Blacklight system at the Pittsburgh Supercomputing Center (PSC).

7. REFERENCES

[1] E. Apostolidis, V. Mezaris, M. Sahuguet, B. Huet, B. Červenková, D. Stein, S. Eickeler, J. L. Redondo Garcia, R. Troncy, and L. Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *MM*, 2014.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[3] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *WWW*, 2012.

[4] P. Das, R. K. Srihari, and J. J. Corso. Translating related words to videos and back through latent topics. In *WSDM*, 2013.

[5] J. Davidson, B. Liebald, J. Liu, et al. The youtube video recommendation system. In *RecSys*, 2010.

[6] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011.

[7] L. Jiang, A. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *MM*, 2012.

[8] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *MM*, 2014.

[9] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. G. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014.

[10] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[11] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.

[12] L. Jiang, W. Tong, D. Meng, and A. G. Hauptmann. Towards efficient learning of optimal spatial bag-of-words representations. In *ICMR*, 2014.

[13] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ICMR*, 2015.

[14] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann. Fast and accurate content-based semantic search in 100m internet videos. In *MM*, 2015.

[15] M. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[16] C. C. Marshall and F. M. Shipman. Saving, reusing, and remixing web video: using attitudes and practices to reveal social norms. In *WWW*, 2013.

[17] Y. Miao, M. Gowayyed, and F. Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. *arXiv preprint arXiv:1507.08240*, 2015.

[18] Y. Miao, L. Jiang, H. Zhang, and F. Metze. Improvements to speaker adaptive training of deep neural networks. In *SLT*, 2014.

[19] Y. Miao, F. Metze, and S. Rawat. Deep maxout networks for low-resource speech recognition. In *ASRU*, 2013.

[20] F. F.-H. Nah. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology*, 23(3):153–163, 2004.

[21] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *MM*, 2004.

[22] R. J. Ordelman, M. Eskevich, R. Aly, B. Huet, and G. Jones. Defining and evaluating video hyperlinking for navigating multimedia archives. In *Companion on WWW*, 2015.

[23] P. Over, G. M. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, and G. Quénot. Trecvid 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics. In *TRECVID*, 2011.

[24] R. Qumsiyeh and Y.-K. Ng. Predicting the ratings of multimedia items for making personalized recommendations. In *SIGIR*, 2012.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 1–42, 2014.

[26] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[27] J. R. Smith. Riding the multimedia big data wave. In *SIGIR*, 2013.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR 2015*, 2015.

[29] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[30] B. Vandersmissen, F. Godin, A. Tomar, W. De Neve, and R. Van de Walle. The rise of mobile and social short-form video: an in-depth measurement study of vine. In *Workshop on Social Multimedia and Storytelling*, volume 1198, pages 1–10, 2014.

[31] B. Varadarajan, G. Toderici, S. Vijayanarasimhan, and A. Natsev. Efficient large scale video classification. *arXiv preprint arXiv:1505.06250*, 2015.

[32] S.-I. Yu, L. Jiang, Z. Xu, et al. Informedia @ trecvid 2014 med and mer. In *TRECVID*, 2014.

[33] S.-I. Yu, L. Jiang, Z. Xu, Y. Yang, and A. G. Hauptmann. Content-based video search over 1 million videos with 1 core in 1 second. In *ICMR*, 2015.

[34] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, 2015.

[35] C.-Z. Zhu and S. Satoh. Large vocabulary quantization for searching instances from videos. In *ICMR*, 2012.