Lookapp – Interactive Construction of Web-based Concept Detectors*

Damian Borth
Dept. of Computer Science,
University of Kaiserslautern
Kaiserslautern, Germany
d_borth@cs.uni-kl.de

Adrian Ulges
German Research Center for
Artificial Intelligence
Kaiserslautern, Germany
adrian.ulges@dfki.de

Thomas M. Breuel
Dept. of Computer Science,
University of Kaiserslautern
Kaiserslautern, Germany
tmb@cs.uni-kl.de

ABSTRACT

While online platforms like YouTube and Flickr do provide massive content for training of visual concept detectors, it remains a difficult challenge to retrieve the right training content from such platforms. In this technical demonstration we present *lookapp*, a system for the interactive construction of web-based concept detectors. It major features are an interactive "concept-to-query" mapping for training data acquisition and an efficient detector construction based on third party cloud computing services.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Search process

General Terms: Algorithms, Human Factors, Performance

Keywords: Web Video, Concept Detection, Query Refinement, Cloud Computing

1. INTRODUCTION

The usage of user-tagged images and videos as training sources for visual learning of semantic concepts gained more attention recently, as it overcomes the need for manual sample annotation [1, 4, 9]. Web-based concept detection systems retrieve training samples for concept learning by formulating a query to platforms like YouTube or Flickr. When formulating such a query a proper mapping between a concept and a set of keywords must be performed and since good queries can be arbitrarily complex – including tags, category restrictions or time/date constraints – a large set of possible configurations for retrieving the right training content exists. This parameter, usually predefined by a human operator, determines strongly the quality of the retrieved training data and therefore the performance of the resulting system: For example, it has been shown for YouTube that to learn a concept like "scenes showing a beach" a refinement to "beach -boys -music" and restriction to the category "Travel & Events" reduces ambiguity and significantly improves the quality of training material [8].

To support the user with the process of such a *concept-to-query* mapping, we present **lookapp***, a system for an interactive construction of web-based concept detectors. The system provides mechanisms to find a proper query formulation for a given concept by offering two key features:

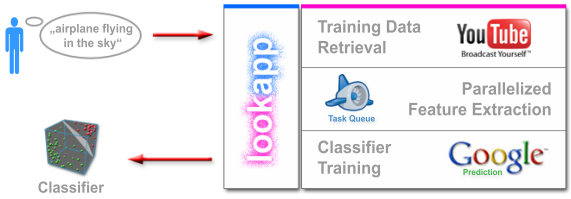

Figure 1: To construct a classifier for a concept like "airplane flying in the sky" the user interacts with the *lookapp* interface to retrieve proper training data and to trigger detector construction.

1. **Interactive Concept-to-Query Mapping**: The core of our system is a retrieval analysis component providing tag and category suggestions. These are based on metadata statistics and the ImageNet Ontology [2] and can be improved further by relevance feedback provided by the user.

2. **Instant Detector Construction**: The second feature is the construction of a concept detector based on the retrieved training data. This is realized using the cloud computing platforms Google AppEngine[1] and Google Prediction API[2]. Additionally, if ground truth is available, immediate performance measurements can be provided about the constructed detectors.

Since video content analysis is a computational intensive task, distributed systems for high-performance computation have been proposed [6, 10], employing parallel programming paradigms on computer clusters, which have to be built up and maintained. In contrast, our system can be deployed on common cloud computing services providing training data acquisition, storage, parallel feature extraction and distributed classifier training.

2. SYSTEM OVERVIEW

An overview of the proposed system can be seen in (**Fig. 1**). The system and its major components are described as following:

*Web demo is available at `http://lookapp.appspot.com`

ICMR'11, April 17–20, 2011, Trento, Italy.
ACM 978-1-4503-0336-1/11/04.

[1]Google AppEngine `http://code.google.com/appengine/`
[2]Google Prediction API: `http://code.google.com/apis/predict/`

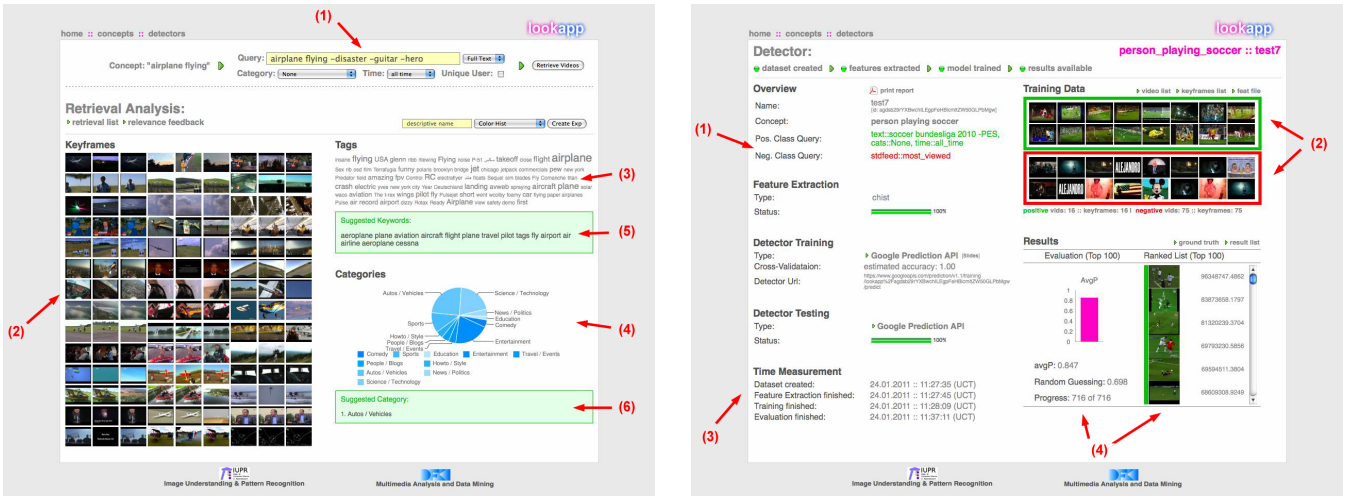

Figure 2: Left: For the current query (1), the performed retrieval analysis displays a mosaic of retrieved keyframes (2), tag occurrences (3) and category distribution (4). Additionally, further tag and category suggestions (5+6) related to the current query are made. **Right**: An overview of the created concept detector, summarizing the used query (1), retrieved datasets (2), time consumption (3) and performance evaluation (4) on a user provided test set.

Training Data Retrieval *Lookapp* uses YouTube as its data source for visual learning. Potential training material is retrieved by querying YouTube via the provided API[3]. Since a semantic concept describes a more complex structure than a simple one-keyword-query could express, an interactive query reformulation and retrieval analysis is offered by the system. The user can add new keywords to the query or exclude keywords to improve retrieval relevance. Training data quality can be further improved by restricting video retrieval to a particular category or employing relevance feedback on keyframe level. A typical retrieval analysis can be seen in (**Fig. 2, left**), where video clips are retrieved for the concept "airplane flying" using the query: "airplane flying -disaster -guitar -hero". The keyframe mosaic already indicates some relevant content for this query. However, the system suggests to further restrict retrieval to the category "Autos / Vehicle" and explore additional keywords like "aviation" or "airport".

Parallel Feature Extraction Videos are represented by keyframes, which are directly fetched from YouTube servers. The system offers color histograms, auto correlograms [3] and SIFT based bag-of-visual-word [5, 7] features, which are extracted for each keyframe.

Classifier Training For classifier training the Google Prediction API is used as an on-demand supervised machine learning service. The Prediction API is build as a distributed system on Google's infrastructure and offers model training in the timespan of minutes. Additionally, if the user provides ground truth for a given concept, the evaluation of the detector is triggered, leading to a ranked list of test keyframes and an average precision (avgP) measurement. The interface showing the building process and performance results of a detector can be seen in (**Fig. 2, right**). As a result either the detector can be used directly by requesting the given Prediction API URL or video and keyframe lists can be downloaded for internal use.

[3]`http://code.google.com/apis/youtube/overview.html`

3. ACKNOWLEDGMENTS

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG), project MOONVID (BR 2517/1-1).

4. REFERENCES

[1] Z. Chen, J. Cao, T. Xia, Y. Song, Y. Zhang, and J. Li. Web Video Retagging. *Multimedia Tools and Applications*, pages 1–30, 2010.

[2] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Int. Conf. Computer Vision and Pattern Recognition*, 2009.

[3] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu, and R. Zabih. Image Indexing Using Color Correlograms. In *Int. Conf. on Pattern Recognition*, 1997.

[4] X. Li, C. Snoek, and M. Worring. Annotating Images by Harnessing Worldwide User-Tagged Photos. In *Int. Conf. on Acoustics, Speech, and Signal Proc.*, 2009.

[5] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[6] F. Seinstra, J. Geusebroek, D. Koelma, C. Snoek, M. Worring, and A. Smeulders. High-Performance Distributed Video Content Analysis with Parallel-Horus. *IEEE Multimedia*, 14(4):64–75, 2007.

[7] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Int. Conf. Computer Vision*, pages 1470–1477, 2003.

[8] A. Ulges. *Visual Concept Learning from User-tagged Web Video*. Phd-thesis, Univ. of Kaiserslautern, 2009.

[9] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning Automatic Concept Detectors from Online Video. *Computer Vision Image Understanding*, 2009.

[10] R. Yan, M.O. Fleury, M. Merler, A. Natsev, and J.R. Smith. Large-Scale Multimedia Semantic Concept Modeling using Robust Subspace Bagging and Mapreduce. In *Proc. Workshop on Large-scale Multimedia Retrieval and Mining*, pages 35–42, 2009.