Automatic Concept-to-Query Mapping for Web-based Concept Detector Training *

Damian Borth
University of Kaiserslautern
D-67663 Kaiserslautern, Germany
d_borth@cs.uni-kl.de

Adrian Ulges
German Research Center for Artificial Intelligence (DFKI)
D-67663 Kaiserslautern, Germany
adrian.ulges@dfki.de

Thomas M. Breuel
University of Kaiserslautern
D-67663 Kaiserslautern, Germany
tmb@cs.uni-kl.de

ABSTRACT

Nowadays, online platforms like YouTube provide massive content for training of visual concept detectors. However, it remains a difficult challenge to retrieve the right training content from such platforms since the underlying query construction can be arbitrarily complex. In this paper we present an approach, which offers an automatic *concept-to-query mapping* for training data acquisition from such platforms. Queries are automatically constructed by a keyword selection and a category assignment using ImageNet and Google Sets as external sources. Our results demonstrate that the proposed method is able to reach retrieval results comparable to queries constructed by humans providing 76% more relevant content for detector training than a one-to-one mapping of concept names to retrieval queries would do.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Indexing

General Terms

Algorithms, Measurement, Experimentation

Keywords

Web Video, Concept Detection, Query Refinement, Query Expansion, Query Mapping

1. INTRODUCTION

As current video collections are growing in size [**?**] the demand for robust search and retrieval tools increases. One successful approach to provide such tools is concept detection [16], which employs supervised machine learning to build an index of semantic concepts for retrieval. Due to the very time consuming acquisition of positive samples for detector training, socially tagged images and video have been considered as a valid alternative to expert labeled training data [8, 20]. Such data is publicly available at large scale from online platforms like Flickr or YouTube and is associated with a noisy but rich corpus of tags, comments and ratings that are provided by their online communities.

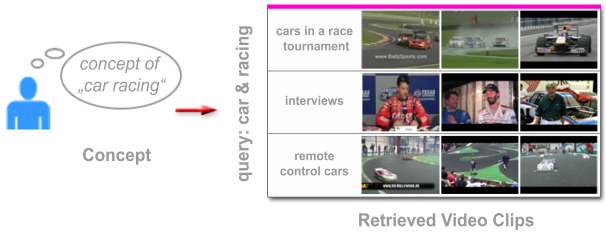

Figure 1: To learn the concept "car racing" a query is formulated to retrieve training material from online portals like YouTube. Unfortunately, a simple query mapping of the concept name will deliver a wide range of related video clips which are not all suitable for visual learning of the concept.

Consequently, prior detector training, such systems must send a query to the desired online platform for training data retrieval. Often, these queries are predefined by a human operator [19] facing the following conditions: First, a proper mapping between a concept definition and a set of keywords is essential as it determines strongly the quality of the retrieved training data and therefore the performance of the resulting system [21]. Second, since good queries can be arbitrarily complex – including tags, category restrictions or time/date constraints – a large set of possible configurations for a query exists, demanding a significant amount of time for query analysis and manual refinement.

This is illustrated in Figure 1, a straightforward mapping of the *concept* "car racing" to the *query* "car racing" may lead to a training set containing non-relevant videos about race driver interviews or clips about remote controlled cars. Knowing this, a query refinement to "car racing tournament -rc -interview" and a restriction to the category "Autos & Vehicle" or "Sports" would reduces ambiguity and increase the amount of relevant content for detector training and finally improve system performance.

*Area chair: Lexing Xie

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

The key contribution of this paper is a novel approach offering such a *concept-to-query mapping*. This is accomplished by two key features:

1. **Automatic Keyword Selection**: The initial query is expanded by additional keywords based on a combination of tag statistics, ImageNet [5] and GoogleSets.

2. **Automatic Category Assignment**: The second feature is the assignment of a category to a query. This is achieved by utilizing the hierarchical structure of ImageNet in combination with tag statistics from YouTube.

Using the proposed approach, we demonstrate that the fraction of relevant content from retrieved training data using automatically constructed queries is comparable to human constructed queries. Also, query construction can be performed on the fly as it stays within the time span of two seconds, which is a tolerable waiting time for web retrieval [11]. Additionally, this functionality was also build into *lookapp*, a public available demo system for construction of web-based concept detectors [4][1].

This paper is organized as follows: first we discuss related work in the context of visual learning and query formulation (Section 2). After this, the proposed approach is presented (Section 3) and evaluated in quantitative experiments on real-world web video data (Section 4). A discussion concludes the paper (Section 5).

2. RELATED WORK

Web data can be considered as an attractive source for visual learning [13, 14, 15, 17, 18, 20] as it allows to train more flexible and scalable visual recognition systems. However, its exploitation is particular challenging due to its subjective and ambiguous nature containing significant amounts of non-relevant material as reported for Google Image Search [14], Flickr images [9] and web video from YouTube [19].

Different approaches have been presented to cope with this challenge. One group of approaches apply relevance learning, where label relevance is modeled either by kernel density estimation [3, 19], or nearest neighbor voting [8]. Another group of methods tackles the problem by tag re-ranking using graph based random walks [6, 10]. An alternative direction of getting more relevant training material can be the improvement of the initial retrieval by directly manipulating the query, an area related to concept-based query expansion or mapping [12, 22, 24]. Here, methods like relevance feedback, lexical approaches including synonyms, hypernyms or statistical approaches including local or global term frequencies and co-occurrences are employed to identify the most relevant concept for a given query. However, in this paper our main goal is not to satisfy users information need during retrieval but to retrieve video material which is suitable for visual learning of concept detectors. Therefore - from our point of view - we aim not to find the most relevant combination of concept detectors for a given query but to find a *proper* query formulation for a given concept definition.

One part of query construction for e.g. YouTube is the assignment of a category. Research in the area of web video categorization was performed based on visual and tag information [2], text and social information [23] and large-scale web crawling and search engine log data [18]. The proposed category assignment is similar to [18], where first tags are recommended and according to this information categories are assigned. However, the setting is different compared to our approach: we select tags and assign categories purely based on concept information and not the visual content of an uploaded video.

[1]Lookapp: http://lookapp.appspot.com

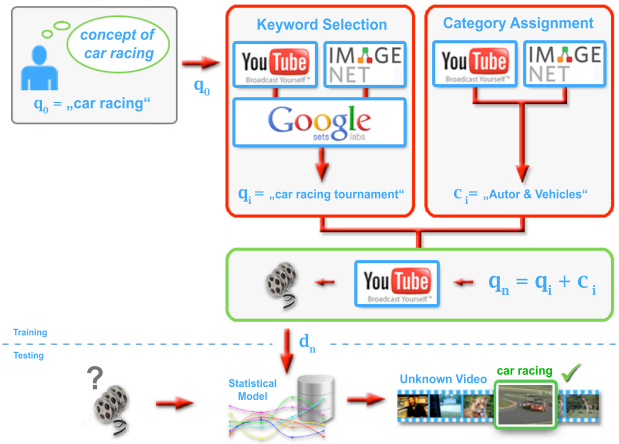

Figure 2: Illustration of the approach. Concepts are trained on downloaded material from online portals like YouTube. To retrieve a list of videos the query q_n must be constructed and send to YouTube (green box). Triggered by an initial concept name q_0 the system performs an automatic keyword selection leading to $q_i = \{car\ racing\ tournament\}$. This keywords are then used to infer a proper category $c_i = \{Autos\ \&\ Vehicles\}$ (red boxes). Both, q_i and c_i, are now used to construct q_n

3. APPROACH

In the following, a framework for query construction in the context of visual learning from the web is described. The system is outlined in Figure 2: to learn a concept like "car racing" video clips are retrieved from YouTube. This is realized by the construction of a query, which is send to the YouTube API. The core of the proposed approach is an automatic keyword selection and category assignment to a given concept. This process is referred to as *concept-to-query mapping* and is highlighted by the red boxes in Figure 2. Taking the original LSCOM [7] concept name as initial query q_0, an automatic selection of keyword terms leads to an expanded query $q_i = \{car\ racing\ tournament\}$. This query is then used to infer a proper category $c_i = \{Autos\ \&\ Vehicles\}$ for the concept. Finally, the query q_n is constructed from q_i and c_i (green box) used to retrieve training data for concept learning.

3.1 Basic Concepts

In the following, a query q represents the set of parameters which is used to retrieve videos from the YouTube API. This set of parameters may including text, tags, category restrictions and limitations to a particular time span or specific country. In this paper we focus on the most distinctive parameter: keywords and categories leading to the query representation $q = \{t_0, ..., t_n\} * \{c_0, ..., c_m\}$ with n keywords t_i and m category assignments c_j. It should be kept in mind

that the presented approach is general as it could be applied to all web video portals that allow access to their database through a similar API like YouTube.

3.2 Automatic Keyword Selection

The first step for the *concept-to-query mapping* is to transfer a concept to a set of keywords. The entry point is given by the concept name forming the initial query q_0. It is important to note that this initial query is expected to retrieve a significant amount of non-relevant videos. Based on q_0 a set of synonyms s_{q_0} is retrieved from ImageNet. Here, ImageNet is preferred over WordNet because it covers concepts suitable for visual learning. Additionally, tag statistics are calculated from the set of videos retrieved by q_0. We calculate tag frequencies for each tag appearing in this initial dataset leading to a ranked list t_{q_0} of top tags for q_0. Although we employ stop word removal and neglect digits and dates, this tag list can be considered noisy and less reliable than s_{q_0} for the purpose of disambiguation of concepts. However, with t_{q_0} we expect to capture specific wording of the YouTube community. Fusing s_{q_0} and t_{q_0} we receive a new query $q_1 = \{s_0, ..., s_{n_s}\} \cup \{t_0, ..., t_{n_t}\} \subseteq s_{q_0} \cup t_{q_0}$ with $n_s + n_t = n$ keywords. This query is now send to Google Sets providing additional semantic relations in the context of q_1. Google Sets is a experimental prototype to generate lists of similar items. Its underlying probability model ranks these items according to their appearance in specific HTML structures as found in the world wide web. As a result we receive a ranked list l_{q_1} of keywords, which we limit to n keywords resulting in the final query $q_i = \{t_0, ..., t_n\} \subseteq l_{q_1}$.

3.3 Automatic Category Assignment

As a second step we automatically assign categories to the previously constructed query q_i. Given q_i, a second set of videos is retrieved from YouTube and its category distribution $p(c|videos)$ is calculated. Additionally, for each keyword $t \in q_i$ its corresponding ImageNet synset is found. If no synset is found for $t \in q_i$, this term will not contribute to the category assignment. For each found synset the path from the synset node to the ImageNet root is build and mapped to a YouTube category according to a manual constructed mapping function $map(p)$. This mapping function maps ImageNet's first (and partially second) level synsets to YouTube categories allowing to transfer all $17k$ synsets to all 15 YouTube categories by only providing roughly 60 manual mappings. A mapping – in this context – may just be as straightforward as $Animal \rightarrow Animals$ or as complex as $University \rightarrow Education$. The final step in the category assignment is a ranking of the mapped YouTube categories according to their query dependent distribution $p(c|videos)$ providing the set $c_i = \{c_0, ..., c_m\}$.

The final query q_n can now be constructed by $q_n = \{t_0, ..., t_n\} * \{c_0, ..., c_m\}$ providing an unambiguous query for training data retrieval.

4. EXPERIMENTS

Experiments are performed on a dataset of real-world web video content retrieved from YouTube. For this, we evaluate the 30 concepts from the TRECVID 2011 benchmark, which have been selected by NIST for evaluation. For each concept three experiments have been performed:

- **[exp-1]** query construction by a simple one-to-one mapping of concept name to a YouTube query. Here, concept names from LSCOM are taken.
- **[exp-2]** query construction by manual refinement from a human according to a visual inspection on YouTube.
- **[exp-3]** query construction performed by the proposed automatic concept-to-query mapping from Section 3. Queries were limited to $n = 3$ keywords and $m = 1$ category assignments.

For each query and experiment 100 videos have been retrieved from YouTube and manually reviewed for relevance according to the LSCOM concept definition. This manual inspection – which was based on three keyframes per video clip – evaluates how many of the retrieved video clips truly contain the concept. Since it has been shown that this fraction $x \leq 1.0$ directly influences detector performance [19], it serves as a metric for the quality of the constructed query.

Table 1 illustrates the results of the evaluation. For each concept the fraction $x \leq 1.0$ of relevant content is shown and additionally for **[exp-3]** the automatically constructed queries are printed (for a full overview of used queries including the manually constructed ones, please refer to the website: http://lookapp.appspot.com/evaluation_mm11). When comparing these three experiments, we can see that **[exp-1]** queries perform weak i.e. they contain the most non-relevant content when retrieving videos from YouTube. Further, manual refined queries **[exp-2]** and automatically constructed queries **[exp-3]** perform comparable to each other improving the fraction of relevant content by 76%.

For some concepts like "airplane flying" or "boat ship" the approach particularly benefits from ImageNet synonyms whereas for concepts, where no synonyms could be found the focus on frequent YouTube tags may lead query construction into the wrong direction like observed for the concept "throwing" or "singing". Also, for concept with a uncommon concept name like "female human face closeup" the method was not able to retrieve any content from YouTube. However, for the majority of concepts the selected keyword terms were semantically meaningful and related to the concept. Also, for most category assignments the method selected the same category a human operator would do.

5. DISCUSSION

In this paper, we have addressed the challenge of automatically mapping concepts to queries for training data retrieval from sources like YouTube. To achieve this we utilize differed external sources like YouTube, ImageNet and Google Sets. In quantitative experiments with web video retrieved from YouTube, it was shown that the proposed approach is able to achieve comparable results to human constructed queries improving the fraction of relevant material by 76% as compared to a simple one-to-one mapping of a concept name to a YouTube query. Future work will extend the approach by relevance feedback to further improve training data retrieval.

6. ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG), project MOONVID (BR 2517/1-1).

Table 1: Results of the evaluation. Fractions of relevant material i.e fraction of videos containing the concept are displayed for each concept and each of the three experiments. The last line displays the relevance as average over all 30 concepts.

Concept Name	[exp-1]	[exp-2]	[exp-3]	[exp-3] Queries (keywords; category)
airplane flying	0.30	0.47	**0.56**	airplane flying aircraft; Autos/Vehicles
animal	0.40	0.89	**0.93**	animal nature; Pets/Animals
Asian people	0.36	**0.52**	0.40	asian people asians; People/Blog
bicycling	0.29	**0.63**	0.62	bicycling city; Sport
boat ship	0.30	0.57	**0.68**	boat ship water; Autos/Vehicles
bus	0.15	0.57	**0.72**	bus buses; Autos/Vehicle
car racing	0.48	0.50	**0.64**	car racing cars; Autos/Vehicles
cheering	0.48	0.27	**0.54**	cheering cheer; Sports
cityscape	0.13	0.12	**0.14**	cityscape architecture; Travel/Events
classroom	0.13	**0.39**	0.36	classroom students; Education
dancing	0.53	0.56	**0.61**	dancing live; None
dark-skinned people	0.62	0.65	**0.79**	dark skinned people; People/Blogs
demonstration or protest	**0.76**	0.72	0.41	demonstration protest funny; News/Politics
doorway	0.07	**0.20**	0.15	doorway vent; Howto/Style
explosion fire	0.33	0.35	**0.61**	explosion fire gasoline; How/Style
female human face closeup	0.04	**0.64**	0.36	female human face closeup; None
flowers	0.11	**0.53**	0.42	flowers green; Howto/Style
ground vehicle	0.31	0.47	**0.70**	ground vehicle military; Autos/Vehicles
hand	0.19	0.51	**0.59**	hand; Science/Technology
mountain	0.11	0.61	**0.70**	mountain peak; Travel/Events
nighttime	0.05	0.28	**0.53**	nighttime building; Travel/Events
old people	0.23	0.23	**0.36**	old people; Comedy
running	0.28	0.35	**0.63**	running basketball; Sports
singing	0.78	**0.88**	0.63	singing fun; None
sitting down	0.02	**0.10**	0.06	sitting down the; Travel/Events
swimming	0.48	**0.78**	0.70	swimming water; Sport
telephones	0.04	**0.67**	0.46	telephone call; Science/Technology
throwing	0.13	**0.65**	0.16	throwing to;None
vehicle	0.31	**0.64**	**0.64**	vehicle car; Autos/Vehicles
walking	0.20	**0.44**	0.09	walking alternative; None
average	0.29	**0.51**	**0.51**	

7. REFERENCES

[1] Thanks, YouTube community, for two BIG gifts on our sixth birthday!. The YouTube Blog; available from http://youtube-global.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html (retrieved: June'11), 2011.

[2] D. Borth, J. Hees, M. Koch, A. Ulges, C. Schulze, T. Breuel, and R. Paredes. TubeFiler – an Automatic Web Video Categorizer. In *Proc. Int. Conf. on Multimedia*, 2009.

[3] D. Borth, A. Ulges, and T.M. Breuel. Relevance Filtering meets Active Learning: Improving web-based Concept Detectors. In *Int. Conf. on Multimedia Information Retrieval (MIR)*, 2010.

[4] D. Borth, A. Ulges, and T.M. Breuel. Lookapp - Interactive Construction of web-based Concept Detectors. In *Int. Conf. on Multimedia Retrieval (ICMR)*, 2011.

[5] 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 (CVPR)*, 2009.

[6] W.H. Hsu, L.S. Kennedy, and S.F. Chang. Reranking Methods for Visual Search. *IEEE Transactions on Multimedia*, 14(3):14–22, 2007.

[7] L. Kennedy, A. Hauptmann, M. Naphade, J. Smith, and S.-F. Chang. LSCOM Lexicon Definitions and Annotations Version 1.0. Technical report, Columbia University, 2006.

[8] X. Li, C.G.M. Snoek, and M. Worring. Unsupervised Multi-Feature Tag Relevance Learning for Social Image Retrieval. In *Int. Conf. on Image and Video Retrieval (CIVR)*, 2010.

[9] Xirong Li, Cees G. M. Snoek, and Marcel Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.

[10] D. Liu, X.S. Hua, L. Yang, M. Wang, and H.J. Zhang. Tag Ranking. In *Int. Conf. on World Wide Web*, 2009.

[11] 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.

[12] A.P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic Concept-based Query Expansion and Re-ranking for Multimedia Retrieval. In *Int. Conf. on Multimedia*, 2007.

[13] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, and J. Han. VideoMule: a Consensus Learning Approach to Multi-label Classification from Noisy User-Generated Videos. In *Int. Conf. on Multimedia*, 2009.

[14] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting Image Databases from the Web. In *Int. Conf. Computer Vision (ICCV)*, 2007.

[15] A. Setz and C. Snoek. Can Social Tagged Images Aid Concept-Based Video Search? In *Int. Conf. on Multimedia and Expo (ICME)*, 2009.

[16] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[17] Y. Sun, S. Shimada, Y. Taniguchi, and A. Kojima. A Novel Region-based Approach to Visual Concept Modeling using Web Images. In *Int. Conf. on Multimedia*, 2008.

[18] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on YouTube: Tag Recommendation and Category Discovery. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[19] A. Ulges, D. Borth, and T. Breuel. Visual Concept Learning from Weakly Labeled Web Videos. In *Video Search and Mining*. Springer-Verlag, 2009.

[20] A. Ulges, C. Schulze, M. Koch, and T. Breuel. The Challenge of Tagging Online Video. *Comp. Vis. Img. Underst.*, 2009.

[21] A. Ulges. *Visual Concept Learning from User-tagged Web Video*. Phd-thesis, University of Kaiserslautern, 10 2009.

[22] D. Wang, X. Li, J. Li, and B. Zhang. The Importance of Query-Concept-Mapping for Automatic Video Retrieval. In *Int. Conf. on Multimedia*, 2007.

[23] C.W. Ngo X. Wu, W.L. Zhao. Towards Google Challenge: Combining Contextual and Social Information for Web Video Categorization. In *Int. Conf. on Multimedia*, 2009.

[24] R. Yan and A.G. Hauptmann. A Review of Text and Image Retrieval Approaches for Broadcast News Video. *Information Retrieval*, 10(4):445–484, 2007.