ARE CONCEPT DETECTOR LEXICONS EFFECTIVE FOR VIDEO SEARCH?

Cees G.M. Snoek and Marcel Worring

Intelligent Systems Lab Amsterdam, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

{cgmsnoek,worring}@science.uva.nl

http://www.mediamill.nl

ABSTRACT

Until now, systematic studies on the effectiveness of concept detectors for video search have been carried out using less than 20 detectors, or in combination with other retrieval techniques. We investigate whether video search using just large concept detector lexicons is a viable alternative for present day approaches. We demonstrate that increasing the number of concept detectors in a lexicon yields improved video retrieval performance indeed. In addition, we show that combining concept detectors at query time has the potential to boost performance further. We obtain the experimental evidence on the automatic video search task of TRECVID 2005 using 363 machine learned concept detectors.

1. INTRODUCTION

Video search engines emphasize different techniques for effective retrieval. The predominant component of most systems is still based on text. Where commercial video search engines rely on keywords from closed captions, speech transcripts, or social tags; academic prototypes typically let users query an archive containing visual feature values rather than the images. Only recently, concept-based retrieval has taken off as a possible alternative technique for video search. It requires semantic video indexing, which is the process of automatically detecting the presence of a semantic concept, like a *beard* or a *helicopter*, in a video stream. An accepted method in semantic video indexing is to use generic methods that learn a detector from a set of examples. This emphasis on generic indexing has opened up the possibility of moving to larger sets of concept detectors. MediaMill has published a collection of 101 machine-learned detectors [10]. LSCOM is working towards a set of 1000 detectors [6]. Thus research in semantic video indexing has now reached the point where over 100, and soon 1000, concept detectors can be learned in a generic fashion, albeit with mixed performance.

The question arises how effective these detector lexicons are for video retrieval, knowing this varying performance. Others have tried to answer this question, e.g. [1–3, 7]. However, these works use either a rather small lexicon (< 20) of concept detectors [2, 3], or evaluate the potential of concept detectors in combination with traditional video retrieval techniques only [1–3, 7]. In contrast to using machine learned concept detectors, [5] investigates the utility of manually annotated examples on video retrieval performance. Hereby avoiding the fact that concept detectors vary in their performance. To the best of our knowledge no work in literature exists, which investigates the effectiveness of using just a large lexicon (> 200) of concept detectors on video search performance. We formulate two hypotheses that we address in this paper. Our first hypothesis states:

Hypothesis 1 *Increasing the number of concept detectors in a lexicon improves video retrieval accuracy.*

Once a large lexicon of concept detectors is available, a combination of some or all of them could further improve video retrieval performance. There is no evidence, however, that combining concept detectors improves the performance of video retrieval systems. This motivates our second hypothesis:

Hypothesis 2 *Combining concept detectors from a lexicon improves video retrieval accuracy.*

To test the two hypotheses we employ a video search using 363 machine learned concept detectors. We obtain experimental evidence from the automatic search task of the 2005 TRECVID benchmark [8].

We organize the remainder of this paper as follows. In Section 2, we introduce video search using a lexicon of concept detectors. Then we present the experimental setup in which we test our two hypotheses in Section 3. We analyze results in Section 4.

2. CONCEPT LEXICON BASED VIDEO QUERYING

We aim to retrieve from a video archive, composed of n unique shots, the best possible answer set in response to a user information need using just a concept detector lexicon. We now detail concept lexicon based video querying.

An extended version of this paper will appear in *IEEE Trans. Multimedia* [9]. This research is sponsored by the BSIK MultimediaN project.

2.1. Building a Concept Lexicon

When building a concept lexicon, $\Omega = \{\omega_1, \omega_2, \ldots\}$, one starts with specifying a set of concepts together with annotated visual examples in the form of key frames. Once these concept annotations are available, one can learn concept detectors by combining feature extraction with supervised machine learning. For our lexicon we adopt the sets of concept annotations made publicly available as part of the MediaMill Challenge [10] and LSCOM [6]. Concepts in these lexicons are chosen based on extensive analysis of video archive query logs and related to program categories, setting, people, objects, activities, events, and graphics. To assure a sound basis for supervised learning, concepts are added to the combined lexicon only when they contain at least 30 positive annotated instances. If concepts appear in both the MediaMill and LSCOM lexicons, we select the one with the best performance on validation data. Ultimately, this process results in a combined lexicon of 363 concepts.

2.2. Lexicon-based Indexing

Given a feature vector $\vec{x}_i$, part of a shot i, the aim in lexicon-based indexing is to obtain a confidence measure, $p(\omega_j|\vec{x}_i)$, for each concept ω_j in Ω. Here, feature extraction is based on the method described in [10], which is robust across different video data sets while maintaining competitive performance. We first extract a number of color invariant texture features per pixel. Based on these, we label a set of predefined regions in a key frame image with similarity scores for a total of 15 low-level visual region concepts. This yields a vector of 15 elements, where each element represents a similarity score to one of the regional concepts. We vary the size of the predefined regions to obtain a total of 8 concept occurrence vectors that characterize both global and local color-texture information. We concatenate the vectors to yield a 120-dimensional visual feature vector per key frame, $\vec{x}_i$.

For machine learning of concept detectors we adopt the experimental setup proposed in [10]. We divide a data set a priori into a non-overlapping train and validation set. The training set $\mathcal{A}$ contains 70% of the data, and the validation set $\mathcal{B}$ holds the remaining 30%. To obtain the confidence measure $p(\omega_j|\vec{x}_i)$ we use the Support Vector Machine framework. Here we use the LIBSVM implementation with radial basis function and probabilistic output. Classifiers thus trained for ω_j, result in an estimate $p(\omega_j|\vec{x}_i)$. We obtain good parameter settings by performing an iterative search on a large number of combinations. We select the parameters with the best performance after 3-fold cross validation, on set $\mathcal{A}$, resulting in $p^*(\omega_j|\vec{x}_i)$. When identical concepts appear in both the MediaMill and LSCOM lexicon, we select the one with best performance on validation set $\mathcal{B}$. We rank concept detection results based on $p^*(\cdot)$ to allow for concept-based querying.

2.3. Concept-based Querying

The set of concepts in Ω forms the basis for querying. For search topics that are related to available concept detectors in the lexicon, a single detector is a good starting point for retrieval. In case the lexicon contains the concept *smokestack*, all information needs related to *chimneys* benefit from using this detector. In practice, a search topic may contain multiple concepts. In such cases, a combination of some or all of them could further improve video retrieval performance. Various combination methods exists. In information retrieval the linear combination of individual methods is often evaluated as one of the most effective combination methods, see for example [4]. The authors of [11] present a theoretical framework for monotonic and linear combination functions in a video retrieval setting. They argue that a linear combination might be sufficient when fusing a small number of detectors. We therefore adopt a linear combination function, similar to [4], which uses a single combination factor λ for pair-wise combination of two concept detectors, defined as:

$$p_2^*(\omega_1, \omega_2|\vec{x}_i) = \lambda \cdot p^*(\omega_1|\vec{x}_i) + (1-\lambda) \cdot p^*(\omega_2|\vec{x}_i), \quad (1)$$

where $\lambda \in [0, 1]$.

3. EXPERIMENTAL SETUP

For evaluation we use the automatic search task of the 2005 TRECVID benchmark [8]. Rather than aiming for the best possible retrieval result, our goal is to assess the effectiveness of concept detector lexicons on video search. To that end, our experiments focus on the evaluation of retrieval strategies using concept detectors only, given an information need. We first determine the best possible single concept detector for an information need, or topic, given an increasing lexicon of concept detectors. Then, we assess the influence of combining concept detectors by fusing individual detector results. We will now detail the search task and our experiments.

3.1. TRECVID Automatic Video Search Task

The TRECVID 2005 video archive contains 169 hours of video data, with 287 episodes from US, Arabic, and Chinese news sources, recorded during November 2004. The test data collection contains approximately 85 hours of video data. The video archives come accompanied by a common shot segmentation, which serves as the unit for retrieval. The goal of the search task is to satisfy a number of video information needs. Given such a need as input, a video search engine should produce a ranked list of results without human intervention. The 2005 search task contains 24 search topics in total. For each topic we return a ranked list of up to 1000 results. The ground truth for all 24 topics is made available by the TRECVID organizers, and to assess our retrieval methods we use *average precision* (AP), following the standard in TRECVID evaluations [8]. We report the mean average precision (MAP) over

all search topics as an indicator for overall search system performance.

3.2. Experiments

We apply the 363 concept detectors from Section 2.2 on each shot from the TRECVID 2005 test set. We then perform two experiments to test our hypotheses and assess the effectiveness of concept detector lexicons for video search:

Experiment 1 *Assessing the effectiveness of increasing concept detector lexicons for video search.*

In the first experiment, we randomly select a bag of 10 concepts from our lexicon of 363 detectors. We evaluate each detector in the bag against all 24 search topics and determine the one that maximizes AP for each topic. Hence, we determine the upper limit in MAP score obtainable with this bag. In the next iteration, we select a random bag of 20 concept detectors from the thesaurus, and once more the optimal MAP is computed. This process is iterated until all concept detectors have been selected. To reduce the influence of random effects – which may disturb our judgement of increasing lexicon size on video search performance in both a positive and negative manner – we repeat the random selection process 100 times.

Experiment 2 *Assessing the effectiveness of pair-wise concept detector combinations for video search.*

In the second experiment, we assess whether combining individual concept detectors makes sense. Since the quality of individual detectors varies, their combination does not necessarily yield improved performance. To avoid the problem of automatically selecting relevant detectors given a user query, we combine the predetermined best and second best concept detector per query. We employ the pair-wise combination from eq. (1), using all possible linear combinations with steps of 0.1 for λ. This allows us to rank all shots according to $p_2^*(\cdot)$. We term this combination "oracle fusion" as it uses the test set results to select the optimal combination on a per-query basis. We include it to explore the upper limits of performance that can be reached by combining two concept detectors.

4. RESULTS AND ANALYSIS

4.1. Experiment 1: Increasing Concept Detector Lexicons

We summarize the influence of an increasing lexicon of concept detectors on video search performance as a box plot in Fig. 1. We observe from the results that a clear positive correlation exists between the number of concept detectors in the lexicon and video retrieval performance. The box plot also shows that the median is shifted towards the bottom of the box for the first 30 concept detectors, even when the outliers are ignored. This indicates that, on average, performance is low for small lexicons, but some detectors perform very well

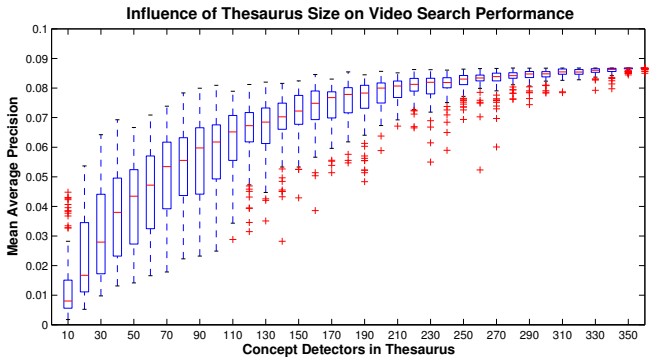

Fig. 1. Box plot showing the positive influence of an increasing lexicon size, in random bags of 10 machine learned concept detectors, on mean average precision over 24 topics from the TRECVID 2005 video retrieval benchmark. Extreme values after 100 repetitions are marked (+) as outliers.

for specific topics. However, it is unlikely that a large variety of topics can be addressed with a small lexicon, which explains the skew. With only 10 randomly selected concept detectors the median MAP score is 0.008. Indeed, the usage of few detectors is of limited use for video retrieval. However, a steady increase in concept detectors has a positive influence on search performance. For the first 60 concept detectors this relation is linear even, increasing MAP from 0.008 to 0.047. When lexicons grow, more search topics can be addressed with good performance. However, the shift towards the high end of the box denotes that a substantial number of concept detectors in our lexicon do not perform accurate enough yet to be decisive for performance. This causes the smoothing effect when more then 70 concept detectors are added. Nevertheless, performance keeps rising until the limit of this lexicon is reached for the maximum obtainable MAP of 0.087. Note that this value is competitive with the state-of-the-art in video search [3, 7, 8].

4.2. Experiment 2: Combining Concept Detectors

The pair-wise oracle fusion of the best and second best concept detector for each of the 24 search topics is summarized in Table 1. The increase in AP for 20 out of 24 search topics indicates that a pair-wise combination of concept detectors pays off in general. It is hard, however, to draw strong conclusion based on analysis of individual search topics. For search topics such as *find shots of a graphic map of Iraq with Bagdhad marked* the pair-wise combination of maps and overlayed text makes sense and simultaneously increases performance. This is not always the case however, yielding questionable detectors as the best choice for some search topics. For the topic *find shots of George Bush entering or leaving a vehicle*, for example, the optimal detectors are *rocket propelled grenades* and *Iyad Allawi*. In this case the pair-wise combination does

Table 1. Best and second best concept detector for each of the 24 TRECVID 2005 search topics, in terms of average precision (AP), together with their pair-wise oracle fusion, where λ indicates the weight from eq. (1). The last column denotes the performance change, over the best concept detector, after fusion.

	Best Concept Detector		**2nd Best Concept Detector**		**Pair-wise Oracle Fusion**		
Search Topic	*Concept Name*	*AP*	*Concept Name*	*AP*	λ	*AP*	*%Change*
Condoleeza Rice	Flag USA	0.024	A. Sharon	0.005	0.1	0.024	0.6
Iyad Allawi	I. Allawi	0.009	A. Sharon	0.004	0.1	0.012	21.1
Omar Karami	Chair	0.028	Meeting	0.022	0.8	0.030	7.0
Hu Jintao	I. Allawi	0.012	H. Nasrallah	0.006	0.6	0.015	18.7
Tony Blair	Election campaign address	0.007	J. Kerry	0.004	0.3	0.008	13.5
Mahmoud Abbas	Conference room	0.013	Meeting	0.008	0.9	0.019	41.6
Graphic map of Iraq, Bagdhad marked	Map	0.027	Overlayed text	0.010	0.2	0.051	89.4
Two visible tennis players on the court	Athlete	0.650	Sports	0.627	0.4	0.673	3.5
People shaking hands	Beards	0.011	Old people	0.005	0.5	0.015	38.1
Helicopter in flight	Helicopters	0.079	Vehicle	0.073	0.9	0.101	27.5
George Bush entering or leaving vehicle	Rocket propelled grenades	0.036	I. Allawi	0.005	1.0	0.036	0.0
Something on fire with flames and smoke	Violence	0.015	Explosion	0.013	1.0	0.015	0.0
People with banners or signs	People marching	0.101	Crowd	0.063	1.0	0.101	0.0
People entering or leaving a building	Muslims	0.004	Animal	0.004	0.3	0.005	14.4
A meeting with a large table and people	Furniture	0.104	Suits	0.067	0.9	0.105	0.8
A ship or boat	Cloud	0.043	Waterscape	0.036	0.7	0.067	56.2
Basketball players on the court	Indoor sports venue	0.280	Dark-skinned people	0.268	0.8	0.290	3.6
One or more palm trees	Weapons	0.023	Walking running	0.018	1.0	0.023	0.0
An airplane taking off	Classroom	0.053	Helicopters	0.046	0.9	0.058	10.2
A road with one or more cars	Car	0.073	Road	0.048	0.6	0.082	13.2
One or more military vehicles	Armored vehicles	0.089	Machine guns	0.065	0.9	0.091	2.2
A tall building	Office building	0.047	Building	0.046	0.8	0.059	25.2
A goal being made in a soccer match	Stadium	0.343	Lawn	0.301	0.9	0.397	15.9
Office setting	Computers	0.009	A. Sharon	0.008	0.1	0.010	1.1
Mean		*0.087*		*0.073*		*0.095*	*9.8*

not match semantically, nor does it yield improved performance. However, for individual topics the increase in AP after combination can be as high as 89%. Overall, the retrieval results increase with almost 10%.

5. CONCLUSION

In this paper, we assess the effectiveness of using a large lexicon of 363 concept detectors on video search performance. We formulate two hypotheses related to the influence of lexicon size and the impact of combining two detectors. Experiment 1 confirms our first hypothesis. It shows that a clear positive correlation exists between the number of available concept detectors and video search performance. Our second hypothesis states that combining concept detectors yields improved video search performance. Our results in experiment 2 seem to confirm this hypothesis. When we combine the predetermined best two concept detectors for a query, the increase is 10% on average. Thus, using a large lexicon of concept detectors for video retrieval is effective indeed. How to automatically select the best possible concept detectors given a user query [9], and how to automatically mix multiple concept detectors that vary in their quality, are open research questions that we aim to address in future work.

6. REFERENCES

[1] S.-F. Chang et al. Columbia university TRECVID-2006 video search and high-level feature extraction. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2006.

[2] M. Christel and A. Hauptmann. The use and utility of high-level semantic features. In *CIVR*, 2005, pp. 134–144.

[3] T.-S. Chua et al. TRECVID 2005 by NUS PRIS. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2005.

[4] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for European languages. In *Proc. ACM SAC*, pp. 1073–1077, 2004.

[5] W.-H. Lin and A. Hauptmann. Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts. In *Proc. IEEE ICME*, Toronto, Canada, 2006.

[6] M. Naphade et al. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.

[7] A. Natsev, M. Naphade, and J. Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proc. ACM Multimedia*, Singapore, 2005.

[8] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. ACM MIR*, Santa Barbara, USA, 2006.

[9] C. G. M. Snoek et al. Adding semantics to detectors for video retrieval *IEEE Trans. Multimedia*, 9(5), 2007. In press.

[10] C.G.M. Snoek et al. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. ACM Multimedia*, Santa Barbara, USA, 2006.

[11] R. Yan and A. Hauptmann. The combination limit in multimedia retrieval. In *Proc. ACM Multimedia*, Berkeley, USA, 2003.