

CAN SOCIAL TAGGED IMAGES AID CONCEPT-BASED VIDEO SEARCH?

Arjan T. Setz and Cees G. M. Snoek

Intelligent Systems Lab Amsterdam, University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands
{asetz, cgmsnoek}@science.uva.nl

ABSTRACT

This paper seeks to unravel whether commonly available social tagged images can be exploited as a training resource for concept-based video search. Since social tags are known to be ambiguous, overly personalized, and often error prone, we place special emphasis on the role of disambiguation. We present a systematic experimental study that evaluates concept detectors based on social tagged images, and their disambiguated versions, in three application scenarios: within-domain, cross-domain, and together with an interacting user. The results indicate that social tagged images can aid concept-based video search indeed, especially after disambiguation and when used in an interactive video retrieval setting. These results open-up interesting avenues for future research.

1. INTRODUCTION

To cater for effective video retrieval, content-based tagging of visual concepts, such as *beach*, *singing*, and *kitchen*, is an important prerequisite. In contrast to video retrieval approaches based on speech transcripts, these methods allow video access on the granularity of the visual semantics. Two visual tagging approaches have become popular: one relies on automatic detection, the other relies on human labeling. While effective to a certain extent, both automatic and manual tagging methods suffer from critical problems.

Automatic concept detection methods rely on invariant visual features in combination with supervised machine learning to train detectors for a wide range of concepts. For each concept detector, labeled examples have to be annotated manually by expert annotators making these annotations expensive and their availability, therefore, limited. Moreover, automatic concept detection is reasonable successful only, as long as the source data on which the detectors are trained are visually consistent with the target data on which they are applied. Cross-domain application of concept detectors is known to be problematic, even after expensive classifier adaptation techniques [10,3]. The limited and domain-specific nature of concept detection training resources prevents scalability to large

amounts of concepts needed for practical and effective video search.

Manual labeling of (broadcast) video has traditionally been the realm of professionals. Because expert labeling is tedious and costly, it typically results in a brief description of a complete video only. In contrast to expert labor, Web 2.0 has launched social tagging, a recent trend to let amateur consumers label, mostly personal, visual content on web sites like YouTube, Flickr, and Facebook. Since the labels were never meant to meet professional standards, amateur labels are known to be ambiguous, overly personalized, and limited per item [5,2]. Manual labeling, whether by experts or amateurs, is geared towards one specific type of use and therefore inadequate to cater for alternative video retrieval needs.

This paper seeks to unravel whether social tagged images can aid concept-based video search. To that end, we present a systematic experimental study that explores the potential of social tagged images as a training resource for automated concept detection. Since the main drawback of using social tagged images is the fact that they are error prone, using social tagged images to feed supervised machine learning potentially translates into deteriorated concept detector performance. Therefore, this experimental study places special interest on the role of disambiguation, where others have only considered the non-disambiguated case [8]. To structure our study, we consider three application scenarios of concept detectors. First, we want to establish whether disambiguating social-tagged images is beneficial when using concept detectors within the social tagged domain itself. We want to establish if, and how, subjectivity influences detector performance before we move on to the cross-domain potential of social-tagged images. For our second, cross-domain, application scenario, we also want to establish the influence of disambiguation on the robustness of the concept detectors. Once cross-domain concept detectors provide an effective first entry in a video collection, a user might be more willing to engage in an interactive session with a video search engine, leading to our third application scenario where we employ detectors trained on social tagged images to kickstart interactive video search. Taken together these three application scenarios cover realistic use-cases to evaluate whether social tagged images can aid concept-based video search.

2. DEFINING THE EXPERIMENTAL STUDY

We study the potential of social tagged images as a training resource for concept-based video search. We do not aim for the best possible performance, but rather focus on the performance gain one can achieve by disambiguating the subjective tags found in social tagged data. We distinguish three application scenarios throughout this paper. For all scenarios we focus on a set of 20 concepts, adopted from the *high level feature extraction* task of the 2008 TRECVID benchmark [6]. These concepts range from objects such as *air-planes* and *boats*, scenes such as *harbors* and *cityscapes* to people-related concepts like *demonstrations* and *drivers*. We consider three datasets throughout this paper and three application scenario experiments, as summarized in Figure 1 and detailed next.

2.1. Datasets

Social tagged images We extract 87K social tagged images and associated labels from the online photo sharing website Flickr using its API. We select the APIs medium setting, which restricts the images to a maximum of 500 pixels for either the height or width of the image

Disambiguated tagged images The set of disambiguated tagged images is the same as the social tagged image dataset, with the exception that the social tags have been disambiguated. Although it might be possible to disambiguate social tagged images automatically [4, 9], such methods cannot guarantee complete and sound disambiguation. Instead, we constructed a disambiguated dataset by manually inspecting the labels of the social tagged image dataset. We eliminate inter-person tagging subjectivity by relying on a single annotator for the entire dataset.

Expert labeled video shots To conduct experiments in a broad domain setting, we use a set of 44K keyframes from the 2008 TRECVID training set [6]: a collection of 100 hours Dutch documentary video. The keyframes all share an uniform image size of 352 by 288 pixels. Collaborative annotation of this set was done by expert annotators around the globe [1]. Similar to the common approach in literature, we assume that annotations for this dataset are sound, complete, and contain minimal tagging subjectivity. Concept annotation statistics for the three datasets are summarized in Table 1.

2.2. Application scenario experiments

We quantify the effectiveness of using social tagged images as training resource for concept-based video search in three application scenarios, each defined as an experiment.

Experiment 1: Within-domain concept detection. To establish the influence of subjectivity found in social tagged images, we compare concept detectors trained on social tagged image data and disambiguated images. We choose the commonly used division of a 67% portion for the training

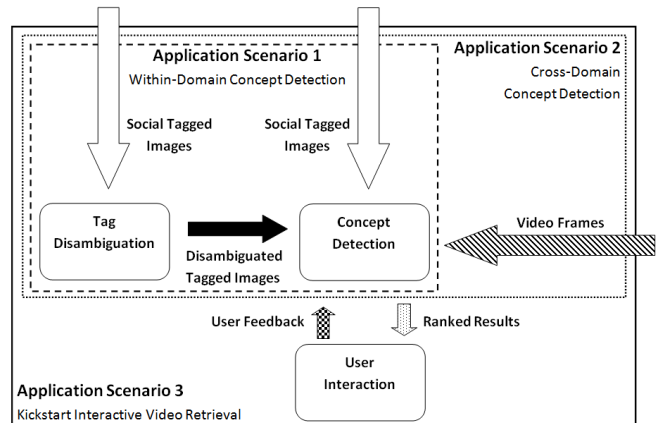


Fig. 1. To establish whether social tags can aid concept-based video search, we consider three application scenario experiments: within-domain concept detection, cross-domain concept detection, and kickstarting interactive video search. For all experiments we compare concept detectors *trained* on social tagged images and disambiguated tagged images.

set and 33% portion for the testing set, per concept. We sample negative examples randomly from the remaining images.

Experiment 2: Cross-domain concept detection. We aim to assess the effectiveness of the detectors trained in experiment 1 when applied to a new domain. In order to do so, we create a laboratory test set for our detectors by using a subset of keyframes from the TRECVID dataset. This dataset consists of all the positive examples in the TRECVID dataset for the concept in question. We supplement these positive images with an equal amount of randomly sampled negative example images.

Experiment 3: Kickstarting interactive video search. Experiment 3 adds (simulated) user interaction to our experiments. We employ the detectors from experiment 1 on the entire TRECVID training set. All ranked results are reviewed to a limited depth n . Since we have a reasonably complete annotation of this set, we are in a position to simulate an interacting user. Both the positive and negative samples of the top n results are added as learning samples for the concept detector and we learn a new cross-domain detector. This process is iterated until no more new correct results are retrieved in the top n .

In order to rule out any random effects from sampling negative examples that may influence the results all three experiments are repeated 100 times.

2.3. Evaluation criteria

For experiments 1 and 2 we evaluate concept detection results using the commonly used average precision measure. For experiment 3 we adopt a different strategy. As we simulate user interaction, relevant keyframes should appear as high as pos-

Table 1. Overview of concept statistics per dataset and experimental results for three application scenarios. *Social* refers to social tagged image training data, *Disam* refers to disambiguated tagged image training data, *Video* refers to expert labeled video data, and *Interaction* refers to average number of user interactions. We report average precision for experiment 1 and 2, and precision at 50 for experiment 3. For all experiments the improvement after disambiguation is indicated by *Gain*. Note that all experiments use a different *test* set.

<i>Concepts</i>	<i>Annotated examples</i>			<i>Experiment 1: Within-domain</i>			<i>Experiment 2: Cross-domain</i>			<i>Experiment 3: Kickstart interaction</i>			
	<i>Social</i>	<i>Disam</i>	<i>Video</i>	<i>Social</i>	<i>Disam</i>	<i>Gain</i>	<i>Social</i>	<i>Disam</i>	<i>Gain</i>	<i>Interactions</i>	<i>Social</i>	<i>Disam</i>	<i>Gain</i>
Airplane Flying	3809	1142	57	0.79	0.91	14.0%	0.93	0.93	-0.1%	7.3	0.04	0.14	258.0%
Boat / Ship	2833	2272	673	0.72	0.74	2.4%	0.80	0.82	2.3%	8.1	0.30	0.33	12.0%
Bridge	2659	1765	232	0.60	0.62	3.3%	0.59	0.63	6.6%	4.4	0.02	0.03	26.0%
Bus	2879	1053	101	0.71	0.78	9.8%	0.72	0.76	4.3%	4.4	0.02	0.04	75.4%
Cityscape	3041	2753	295	0.69	0.76	9.6%	0.74	0.79	6.4%	7.0	0.13	0.21	63.2%
Classroom	2648	1330	342	0.61	0.77	26.2%	0.63	0.66	3.8%	5.3	0.02	0.06	195.0%
Demonstration/protest	3174	2053	258	0.82	0.84	2.6%	0.69	0.70	1.6%	5.4	0.10	0.10	-2.5%
Dog	3874	1488	161	0.64	0.67	4.5%	0.62	0.65	4.6%	2.6	0.00	0.00	75.0%
Driver	3514	2281	494	0.41	0.51	22.6%	0.63	0.69	8.1%	8.5	0.04	0.19	352.9%
Emergency Vehicle	2760	1802	128	0.69	0.80	15.2%	0.77	0.76	-0.7%	6.4	0.03	0.11	231.1%
Flower	3260	2789	764	0.82	0.82	0.5%	0.55	0.55	-0.4%	5.0	0.05	0.04	-28.5%
Hand	2743	2550	2340	0.71	0.76	7.0%	0.68	0.70	2.7%	7.8	0.46	0.70	53.2%
Harbor	2164	1231	263	0.63	0.71	14.0%	0.78	0.80	1.8%	8.1	0.13	0.20	56.5%
Kitchen	3892	2975	395	0.68	0.84	22.4%	0.67	0.70	3.3%	4.3	0.01	0.03	215.2%
Mountain	3533	1170	335	0.78	0.84	8.4%	0.79	0.82	4.0%	7.4	0.19	0.22	16.4%
Nighttime	3396	2580	595	0.73	0.74	1.1%	0.81	0.83	1.5%	8.4	0.26	0.38	48.8%
Singing	3335	1421	555	0.50	0.62	23.0%	0.56	0.60	7.7%	7.9	0.02	0.29	1613.1%
Street	3355	469	2648	0.52	0.62	18.8%	0.59	0.70	14.4%	7.8	0.17	0.73	328.2%
Telephone	3480	847	380	0.40	0.46	13.7%	0.50	0.55	8.6%	2.8	0.00	0.01	1600.0%
Two People	3353	1811	4165	0.43	0.44	4.4%	0.56	0.54	-3.5%	8.8	0.15	0.27	78.0%
<i>Average</i>	–	–	–	0.64	0.71	10.5%	0.68	0.71	3.7%	6.4	0.11	0.20	90.5%

sible in the results. Modern video search engines use approximately 25 keyframes per result page. We assume considerable user incentive and use the equivalent of two pages of retrieval results, yielding *precision at 50*, so $n = 50$.

2.4. Detector implementation

In order to train a semantic concept detector, feature vectors are extracted all datasets. As the effectiveness of features is not the focal point of this paper, we do not rely on the most robust features available. Instead we use 240-dimensional Weibull and Gabor features, which have proven effective in the 2006 MediaMill TRECVID system [7]. For the classifier, we chose the Fisher linear classifier because of its modest computational cost.

3. RESULTS

3.1. Experiment 1: Within-domain concept detection

The results of experiment 1 in Table 1 clearly indicate that within-domain concept detection based on social tags benefits from disambiguation. Results improve for all concepts, the gain ranges from 0.5% to as much as 26.2%, and 10.5% on average. Disambiguating social tagged images seems less effective for visually inconsistent concepts, like *telephone* (Table 2), *driver* and *two people*, as in these cases a robust detector is so much harder to achieve. However, when disambiguated concepts are visually consistent within the do-

main, like *kitchen*, *classroom* and *airplane*, a substantial performance increase becomes possible (Table 2).










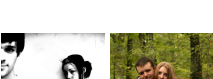
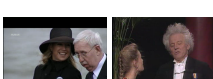

3.2. Experiment 2: Cross-domain concept detection

We summarize the results of experiment 2 in Table 1. As was to be expected, detector performance suffers from the transfer to a new domain. Although the difference between training on social tagged or disambiguated images is less profound, disambiguating still proved beneficial for 16 out of 20 concepts. As anticipated, visual consistency across domains is an important factor. We observe that concepts that are visually inconsistent across domains show poor performance. Moreover, they improve only slightly after disambiguation, consider for example *two people* in Table 2. In contrast, concepts that are visual consistent across domains, benefit from disambiguation, e.g. *street* and *cityscape* in Table 2. These results suggest that cross-domain concept detection benefits from disambiguating social tagged images, especially when concepts are visually consistent across domains.

3.3. Experiment 3: Kickstart interactive video search

The results of experiment 3 in Table 1 indicate that disambiguating also helps when using concept detectors trained on social tagged images to kickstart cross-domain video retrieval. On average the precision at 50 results increase 90.5% from 0.11 to 0.20, with an average of 6.4 user interaction cycles. Continuing the trend of experiment 2, we observe

Table 2. Example concept detection results when using disambiguated tagged images as training resource. For all experiments the first two rows show good performing concepts and the bottom row shows modest performing concepts. Note the (in)consistency in visual appearance of good and modest performing concepts. Experiment 2 highlights images from both the training and test set to illustrate cross-domain visual (in)consistency.

<i>Experiment 1: Within-domain</i>	<i>Experiment 2: Cross-domain</i>		<i>Experiment 3: Kickstart interaction</i>
	<i>Train</i>	<i>Test</i>	
 Classroom	 Street		 Driver
 Kitchen	 Cityscape		 Hand
 Telephone	 Two People		 Flower

that concepts which are visually consistent across domains achieve the best scores. Consider for example *hand* and *street* in Table 2. We further observe that results for visually inconsistent concepts, such as *flower* in Table 2, resemble the visual appearance of the concept in the social tagged image domain. In this example, close-ups of brightly colored flowers. For 18 out of 20 concepts the disambiguated approach yields improved performance, with only few user interactions, in some cases, the increase in performance is substantial.

4. CONCLUSION

In this paper we investigate whether social tagged images can act as a training resource for concept-based video search. In particular, we explore the effect that subjectivity found in social tags has on concept detector performance. We do so by comparing two classifiers, trained on social tagged images and disambiguated tagged images respectively, under three different application scenarios. For within-domain application, our results show that disambiguation increases detector performance for all concepts, with an average gain of 10.5%. When applied across domains, performance deteriorates, but disambiguation of social tagged images still aids cross-domain concept detection for 16 out of 20 concepts. When the cross-domain detectors are used to kickstart interactive video retrieval the disambiguated detectors reveal their biggest potential. In a (simulated) active learning-like scenario we report an average performance gain of 90.5% over detectors trained on non-disambiguated social tagged images. The results of our experimental study suggest that after disambiguation, social tagged images can be a valuable aid for concept-based video search, especially when used in interaction with the user. They are not a viable alternative to expert

annotated training resources yet, but the results do suggest promising avenues for future research.

5. REFERENCES

- [1] S. Ayache and G. Quénot. Evaluation of active learning strategies for video indexing. *Image Communication*, 22(7-8):692–704, 2007.
- [2] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
- [3] W. Jiang, E. Zavesky, S.-F. Chang, and A. C. Loui. Cross-domain learning methods for high-level visual concept classification. In *Proc. IEEE ICIP*, San Diego, USA, 2008.
- [4] X. Li, C. G. M. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proc. ACM MIR Conference*, Vancouver, Canada, 2008.
- [5] K. K. Matusiak. Towards user-centered indexing in digital image collections. *OCLC Systems & Services*, 22(2):263–296, 2006.
- [6] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *Proc. ACMMIR Workshop*, 2006.
- [7] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2006 semantic video search engine, In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2006.
- [8] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proc. ACM CIVR*, Niagara Falls, Canada, 2008.
- [9] K. Q. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *Proc. ACM Multimedia*, Vancouver, Canada, 2008.
- [10] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proc. ACM Multimedia*, Augsburg, Germany, 2007.