Evaluating Sources and Strategies for Learning Video Concepts from Social Media

Svetlana Kordumova
Intelligent Systems Lab Amsterdam
University of Amsterdam
The Netherlands
Email: s.kordumova@uva.nl

Xirong Li
MOE Key Lab of DEKE
Renmin University of China
China
Email: xirong@ruc.edu.cn

Cees G.M. Snoek
Intelligent Systems Lab Amsterdam
University of Amsterdam
The Netherlands
Email: cgmsnoek@uva.nl

***Abstract*—Learning video concept detectors from social media sources, such as Flickr images and YouTube videos, has the potential to address a wide variety of concept queries for video search. While the potential has been recognized by many, and progress on the topic has been impressive, we argue that two key questions, i.e., *What visual tagging source is most suited for selecting positive training examples to learn video concepts?* and *What strategy should be used for selecting positive examples from tagged sources?*, remain open. As an initial attempt to answer the two questions, we conduct an experimental study using a video search engine which is capable of learning concept detectors from social media, be it socially tagged videos or socially tagged images. Within the video search engine we investigate six strategies of positive examples selection. The performance is evaluated on the challenging TRECVID benchmark 2011 with 400 hours of Internet videos. The new experiments lead to novel and nontrivial findings: (1) tagged images are a better source for learning video concepts from the web, (2) selecting tag relevant examples as positives for learning video concepts is always beneficial and it can be done automatically and (3) the best source and strategy compare favorably against several present-day methods.**

I. Introduction

Many videos are produced every day. In order to pinpoint arbitrary fragments of a video with respect to a specific query, semantic labels at a shot or even frame level are prerequisites. Consider for example the amount of frames that are contained in the 72 hours of Internet videos that are being uploaded to YouTube every minute. This leaves us no other choice but to devise machine tagging mechanisms that can detect visual concepts such as *animal*, *building*, and *snow* at the frame level [2], [13]. The state-of-the-art in video concept detection is to learn SVM classifiers from manually labeled frames represented by visual code features [3], [17]. However, the expense of manual labeling results in training examples with limited availability. As a consequence, the performance of concept detection is bounded to a narrow application domain where a limited array of concepts can be reliably detected [19].

In order to detect all possible video concepts one can think of, a promising line of research is to automatically acquire training examples from social media, where many socially tagged videos and images exist. Ulges *et al.* were among the first to learn video concepts from YouTube [16]. In their system, if social tags of a video match a given concept, all frames of the video are used as positive examples of that concept. Setz and Snoek [12] conducted a pilot study on learning video concepts from Flickr images, directly treating images labeled with the concept as positives. However, social tags are known to be unreliable and often irrelevant with respect to the visual content they are describing [5], [7], [8], [15], [18], [20]. Hence, for learning meaningful concept detectors from social media, selecting appropriate examples from a proper data source is crucial.

Fig. 1. What visual tagging source is most suited for learning video concept detectors?

As noted by Yang *et al.* [19], the performance of a concept detector could degenerate severely if the training and the test videos are from different genres, *e.g.*, broadcast news and documentaries. However whether this will also hold true in a cross-source scenario, say applying image classifiers on video data, has not been investigated yet. We observe prevailing usage of socially tagged videos as the training source of choice [4], [15], [16], [18], even though selecting positive training examples from videos is more difficult than selecting positive examples from images. This is not only because video tags are noisy, but also a question on how to propagate tags to the frame level is still open [1], [4]. However whether propagating the tag to the frame level is more beneficial then simply using the tagged images has not been addressed yet. The research question thus arises: *What visual tagging source is most suited for selecting positive training examples for learning video concept detectors*?, see Figure 1.

To acquire accurate positive examples from social media, a common approach is through a retrieval process, where socially tagged examples are first ranked in terms of their estimated relevance scores with respect to a given concept [7], [14], [15], [20]. Then, the top ranked proportion of the examples is preserved. In [7], [20], for instance, a fixed number of examples are selected for every concept, while [15] just tried a varying amount of frame fractions. Some even ignored the unreliability of the social labels, and directly used tagged videos [16] or tagged images [12] as positive examples. So the second question arises: *What strategy should be used for selecting positive examples from tagged sources?*.

We investigate in this paper what tagged sources and what strategies are most suited for selecting positive training examples to learn video concept detectors. We structure our paper as an experimental study with a complete system that learns concepts from social media. Our three main contributions are: First, we systematically compare socially tagged videos and images as training data for video concept

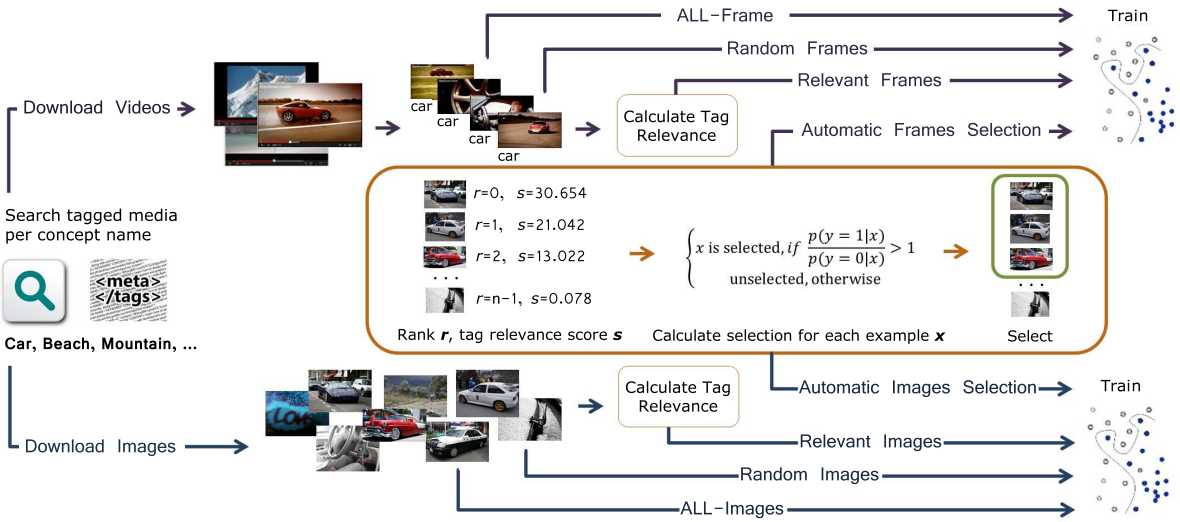

Fig. 2. We perform an experimental study with a video concept search engine to evaluate what source and strategy are most suited for learning video concept detectors from social media.

detection, resulting in new and non-intuitive findings. To the best of our knowledge, such a comparison has not been done before. Second, we compare six strategies for positive example selection. We show that an automatic selection strategy of relevant positive examples is possible, reaching a near-optimal selection. For now we rely on a simple selection strategy, which in the future can be improved with smarter sampling or more dedicated machine learning methods. Third, the best source and strategy outperform several present day alternatives on the challenging TRECVID benchmark.

II. Related Work

In a representative work by Ulges *et al.* [16], video concept detectors are trained using YouTube videos by directly treating social tags as relevant labels for all video frames. Wang *et al.* [18] follow a similar approach, but use a set of manually labeled videos to bootstrap the learning process. Setz and Snoek [12] investigate whether Flickr images can be exploited as a direct training resource for learning video concepts. Notice that none of the above works compares which *source* is a better choice. This study covers the approaches of Ulges *et al.* [16] and Setz and Snoek [12], naming them as the ALL-FRAME and the ALL-IMAGE baselines respectively.

For video it is easy to observe that even when tags are relevant at the video level, it does not imply that the tags are relevant for all shots and frames as well. Ulges *et al.* [15] model the relevance of video (key)frames with density estimation in the feature space. Thus, if a video is labelled with multiple tags, their method would not find representative frames for all tags separately, since their density estimation operates on the feature space, not the tag space.

Positive example estimation of tagged images has been addressed by many using the semantic field of the user provided tags [20], or by neighbor voting of visual neighbors [5]. The semantic field method [20] determines tag relevance in terms of tag-wise similarity. Since the semantic field considers the tags only, it cannot identify relevant frames for individual tags. Li *et al.* [5] learn the relevance of tags accompanying an image with a neighbor voting algorithm. They exploit the observation that similar tags issued for similar images are reliable, by accumulating tag votes over visually similar images. Including a diverse set of visual features in the neighbor voting algorithm further improves the effectiveness of tag relevance [6]. An alternative to tag relevance is proposed by Liu *et al.* [8]. Their method is also founded on neighbor voting, but the neighbors are weighted with a Gaussian function. Since [8] restricts the neighbors to be images labeled with the tag, while [5] exploits the entire collection, we consider the latter better suited for positive example selection.

In the literature, the top ranked proportion of the examples is preserved [7], [20], or multiple fractions are evaluated and the best one is determined at the expense of annotation labor [15]. Li *et al.* [4] bypass the selection problem by relying on Multiple Instance Learning (MIL) [11] to estimate the relevance of a video tag at shot level. In a MIL setting, instances are organized into bags and it is the bags, instead of individual instances that are labeled for training. In [4] the tagged videos are considered as labeled bags, and the frames are the instances. If a video is labeled with a concept, all its frames are considered as positive. However, the base MIL model [11] will hurt if videos are miss-labelled, and it is known that the social tags are imperfect and ambiguous. To overcome this obstacle, [4] calculates a tag correctness score and uses it as a weight in the optimization function of the MI logistic regression. For a test frame, its predicted score is further smoothed using scores of its temporal neighbors. However it has been shown that for content based image retrieval, supervised learning is superior over MIL [10]. This study evaluates [11] and [4], naming them as MIL and MIL+ respectively. We investigate whether surpassing the selection problem, like is done in MIL, is a better strategy than selecting the positive examples before training a concept detector.

III. Experimental Video Search Engine

To answer the research questions raised in the introduction, we structure our paper as an experimental study with a complete system that learns concepts from social media. We identify three key components of such a general system, A) harvesting social media sources, B) diverse strategies for selection of positive training examples and C) concept detection using state-of-the-art implementation *e.g.,* [3], [17]. The dataflow of our system is highlighted in Figure 2.

TABLE I. POSITIVE EXAMPLES FOR 20 CONCEPTS IN BOTH THE TAGGED VIDEO AND TAGGED IMAGES DATASETS.

Concept	YouTube Videos	Video Frames	Flickr Images
Animal	191	7506	7506
Beach	185	6554	6554
Building	197	5207	5207
Car	192	8508	8508
Child	174	6938	6938
City	169	8326	8326
Face	169	7210	7210
Hand	179	5835	5835
Landscape	181	4005	4005
Mountain	181	6445	6445
Oceans	150	5755	5755
Outdoor	184	6103	6103
Plant	151	5139	5139
Road	170	8047	8047
Sky	154	6261	6261
Snow	177	7467	7467
Sports	196	6685	6685
Streets	141	7493	7493
Trees	158	5805	5805
Vehicle	184	4621	4621

A. Harvesting social media sources

We harvest two type of media sources, tagged videos and tagged images. The tagged videos are collected from YouTube, which is one of the most popular service for video sharing. The tagged images are selected from Flickr as one of the most popular sharing service for images. We learn detectors for 20 concepts, covering objects like *Car, Plant* and scenes like *Outdoor, Landscape*. We name the two sources as *Tagged Videos* and *Tagged Images*.

Tagged Videos. In order to construct a diverse set, we collect videos retrieved by four distinct ranking criteria, i.e., view count, relevance, date published and user rating. We obtain for each of the 20 concepts four lists of retrieved videos and their metadata, containing videos id, tags, author, video duration etc. The date published ensures that new instances of concepts like *Car* and *Building* models are covered. We include in our dataset the top 50 retrieved videos from each of the four ordering criteria. Hence, for each concept we download the most viewed, most relevant, most recently uploaded and best rated videos. We shot segment each video and define the middle frame of each shot as a keyframe. Since we want to show the influence of frame selection, we maintain only those videos that have at least two shots. This process resulted in 200 hours of web video and 130K keyframes.

Tagged Images. We adopt the Flickr image collection from [5]. The images are of medium size with width or height fixed to 500 pixels. A subset from this dataset is selected as training source for videos. Considering fair comparison, the number of images for each of the 20 concepts we maintain the same as the number of frames from the *Tagged Videos* dataset. In total, this collection has 130K images.

The number of examples per concept for both sources are shown in Table I.

B. Selection strategies

Given a specific concept ω, we select positive training examples from the two sources described in Section III-A. Let x be such an example in consideration. Depending on the source of training data, x is either an image labeled with ω or a frame extracted from a video labeled with ω. One selection strategy is to rely on the social tags and to randomly select tagged examples x with the concept ω. However, as aforementioned, due to the subjective nature of social tagging, simply treating x as a positive example may be problematic. Therefore another strategy is to calculate tag relevance per example, and then select the top ranked examples. For every example x labeled with concept ω, we use the multi-feature variant of the neighbor voting algorithm [6] to compute tag relevance scores.

The tag relevance scores allow us to rank the examples such that the most relevant examples are deemed to be placed at the top. As noted earlier, how many of the top examples should be selected remains a question. Instead of setting an ad hoc threshold, in this work we make an endeavor to determine a cut-off automatically. To this end, we consider a simple strategy, calculating which example x should be selected based on some decision rule. We employ the Bayesian decision rule because of its effectiveness. For each example, we use s to denote its tag relevance score and r to denote the corresponding rank, where $r = 0, \ldots, n-1$, and n is the number of examples labeled with ω.

We observe that the first tag relevant example is very often a positive response, see Figure 3. Hence, we use this first-hit example as a reference point to estimate the probability of the other examples being positive. In that regard, we introduce a binary random variable y, where $y = 1$ means x is positive, and 0 otherwise. The problem of positive example selection boils down to estimating the conditional probability $p(y = 1|x)$. With the Bayesian decision theorem, we define the selection rule simply as:

$$\begin{cases} \text{x is selected, if } \dfrac{p(y=1|x)}{p(y=0|x)} > 1, \\ \text{unselected, otherwise.} \end{cases} \tag{1}$$

For computing $p(y = 1|x)$, we have access to two observations with respect to x, *i.e.*, the relevance score s, and the corresponding rank r. Using a single observation is limited, since the scores are discriminative but less robust, while the quantized ranks tend to be more robust but less discriminative. Hence, we consider their combination, which should result in a better estimation of $p(y = 1|x)$. Using probability algebra, we have $p(y|x) = p(y|s, r) = p(s, r|y)p(y)/p(s, r)$. For $p(s, r|y)$, we make a practical simplification by estimating it through $p(s|y) \cdot p(r|y)$. We also expand $p(s|y)$ and $p(r|y)$ using Bayes' theorem. Accordingly, we rewrite

$$p(y|x) \approx \frac{p(y|s)p(y|r)p(s)p(r)}{p(s, r)p(y)}, \tag{2}$$

For the unknown prior $p(y)$, an uniform prior assumption is reasonable, and the decision function is

$$\frac{p(y=1|x)}{p(y=0|x)} = \frac{p(y=1|s)p(y=1|r)}{p(y=0|s)p(y=0|r)}. \tag{3}$$

With the intuition that examples with larger tag relevance scores and higher ranks are more likely to be positive, we simply approximate $p(y = 1|s)$ as

$$p(y=1|s) \approx \frac{s}{s_{max}}, \tag{4}$$

where s_{max} is the score of the top ranked example, and compute $p(y = 1|r)$ as

$$p(y=1|r) \approx 1 - \frac{r}{n}. \tag{5}$$

C. Concept detection

We follow a state-of-the-art bag of visual codes pipeline to train video concept detectors [17]. We compute SIFT, OpponentSIFT and

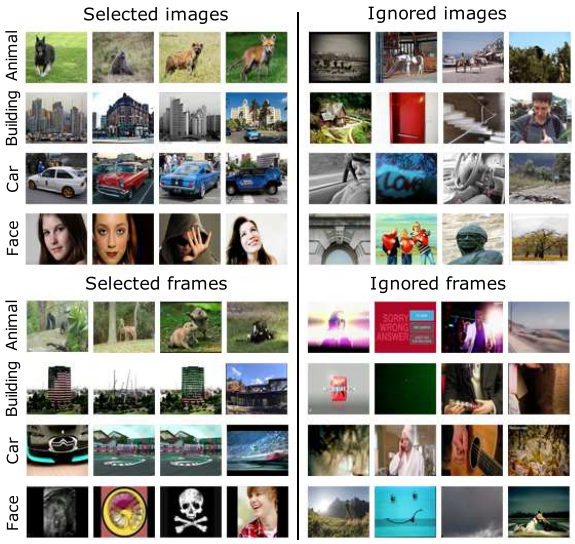

Fig. 3. Positive training examples automatically selected from social media by this paper. Left columns show the selected images and frames, while the right columns show images and videos labeled with a given concept but discarded by the selection strategy for relevant examples.

RGBSIFT descriptors at (1) Harris-Laplace keypoints and (2) dense sampled points, at every 6 pixels for two scales. As visual features we employ a spatial pyramid of 1x1 and 1x3. The codebook size is 4,096, constructed with k-means clustering. As classifier we employ a one-vs-all Support Vector Machines with the fast histogram intersection kernel [9] for its high efficiency. The SVM models are optimized by 3-fold cross validation.

IV. Three Experiments

A. Test Set

As test data we adopt the challenging internet video collection from the TRECVID 2011 benchmark [13]. We use the development dataset provided for the Semantic Indexing task, which consists of 400 hours of Internet Archive videos, having 263,355 shots. Each shot is represented with a single keyframe. The dataset comes with ground truth annotations on a keyframe level, including the 20 concepts identified in our experiments. We evaluate all runs on this set.

B. Experiment 1: What source?

In experiment 1 we address the research question *what visual tagging source is most suited for selecting positive training examples to learn video concepts?*. We use the *Tagged Videos* and the *Tagged Images* described in Section III-A as instantiations of two distinct visual tagging sources. We learn concept detectors using the ALL-FRAME [21] and the ALL-IMAGE [15] scenarios separately, and compare their performance.

C. Experiment 2: What strategy?

Besides the ALL-FRAME and the ALL-IMAGE baselines, we introduce the following six additional example selection strategies, where the first three are based on the *Tagged Videos* and the last three are based on the *Tagged Images*.

Strategy 1a. Relevant Frames. We calculate a tag relevance score for each keyframe before ranking. A fraction of the top ranked frames for each concept are selected as positive training examples. We vary the fraction to investigate the influence of a varying number of positive examples.

Strategy 1b. Random Frames. We randomly sample positive frames from the *Tagged Videos* for each concept. We vary the number of sampled positive examples to investigate their influence.

Strategy 1c. Automatic Frame Selection. We rely on the same frame ranks as in strategy 1a, but here we estimate the selection with the simple Bayes approach described in section III-B.

Strategy 2a. Relevant Images. We calculate a tag relevance score for each image before ranking. A fraction of the top ranked images for each concept are selected as positive training examples. We vary the fraction to investigate the influence of a varying number of positive examples.

Strategy 2b. Random Images. We randomly select images from the *Tagged Images* as positive training data per concept. We vary the number of selected positive examples to investigate their influence.

Strategy 2c. Automatic Images Selection. For this strategy we rely on the images ranked by tag relevance per concept. Due to the large variance of the tag relevance scores for *Tagged Images*, we smooth the scores using the common logarithm. We estimate the selection with the simple Bayes approach described in section III-B.

We would like to stress that the same set of negatives were used for all video strategies 1a, 1b, 1c and the All-FRAME baseline. For each concept we sample the negatives randomly from the other concepts and limit the number to five times the number of positive examples. For the image strategies 2a, 2b, 2c and the All-IMAGE baseline we follow a similar protocol, but here the negatives are sampled from images. While we are aware that selecting appropriate negative examples per concept benefits generalization [7], we prefer to keep the set of negatives constant in our experiments so that we can properly evaluate the influence of selecting positive examples. We refer to Table I for the numbers.

D. Experiment 3: Present-day comparison

We compare the sources and strategies with four present day methods we discussed in section II.: Multiple-instance learning (MIL) [11], Context aware multiple-instance learning with temporal smoothing (MIL+) [4], ALL-FRAME [16] and ALL-IMAGE [12]. For MIL

TABLE II. **Experiment 1: What Source?** Answer: socially tagged images.

Concept	Tagged Videos	Tagged Images
Animal	0.053	**0.088**
Beach	0.278	**0.386**
Building	0.343	**0.496**
Car	0.164	**0.249**
Child	0.048	**0.104**
City	0.067	**0.137**
Face	0.359	**0.660**
Hand	0.098	**0.169**
Landscape	0.317	**0.539**
Mountain	0.079	**0.495**
Oceans	0.106	**0.472**
Outdoor	**0.747**	0.696
Plant	0.171	**0.198**
Road	0.193	**0.358**
Sky	0.311	**0.554**
Snow	0.092	**0.289**
Sports	0.119	**0.127**
Streets	0.119	**0.202**
Trees	0.490	**0.704**
Vehicle	0.225	**0.282**
mAP	0.219	**0.360**

TABLE III. **EXPERIMENT 2: WHAT STRATEGY?** ANSWER: USING AUTOMATIC RELEVANCE SELECTION OF POSITIVE EXAMPLES FROM VISUAL TAGGING SOURCES.

	Strategies using tagged videos				Strategies using tagged images			
Concept	1a. Lower-bound	1a. Upper-bound	1b. Random	1c. Automatic	2a. Lower-bound	2a. Upper-bound	2b. Random	2c. Automatic
Animal	0.059	0.085	0.039	0.081	0.100	0.110	0.092	**0.110**
Beach	0.278	0.267	0.262	0.276	0.301	0.358	0.360	**0.359**
Building	0.350	0.372	0.181	0.379	0.457	**0.532**	0.475	0.531
Car	0.174	0.240	0.150	0.239	0.299	**0.302**	0.232	0.296
Child	0.052	0.068	0.042	0.060	0.099	0.108	0.099	**0.111**
City	0.075	0.136	0.061	0.103	0.147	0.156	0.121	**0.158**
Face	0.465	0.542	0.366	0.504	0.592	0.650	0.645	**0.651**
Hand	0.101	0.102	0.099	0.102	0.160	**0.172**	0.170	0.171
Landscape	0.328	0.466	0.234	0.427	0.474	0.538	0.525	**0.539**
Mountain	0.092	0.292	0.043	0.193	0.440	**0.554**	0.553	0.553
Oceans	0.113	0.383	0.183	0.353	0.468	0.496	0.486	**0.490**
Outdoor	0.763	0.777	0.800	**0.824**	0.725	0.754	0.677	0.737
Plant	0.174	0.273	0.155	**0.263**	0.164	0.185	0.195	0.187
Road	0.207	0.243	0.149	0.242	0.311	**0.336**	0.331	**0.336**
Sky	0.360	0.378	0.230	0.423	0.447	**0.528**	0.501	0.516
Snow	0.073	0.209	0.068	0.217	0.239	**0.318**	0.282	0.308
Sports	0.129	0.152	0.125	0.145	0.188	**0.190**	0.130	0.186
Streets	0.124	0.150	0.102	0.150	0.210	**0.224**	0.179	0.215
Trees	0.504	0.553	0.492	0.548	0.599	0.677	0.669	**0.681**
Vehicle	0.194	0.281	0.162	0.313	0.307	**0.316**	0.272	0.314
mAP	0.231	0.298	0.197	0.292	0.336	0.375	0.350	**0.372**

we use the publicly available code of [11]. The first three approaches rely on the *Tagged Videos*, whereas the last approach relies on the *Tagged Images*. We report results on the *Test Set*.

Evaluation criteria. We adopt average precision, a common approach in the video retrieval literature [13] and mean average precision to evaluate the overall performance.

V. RESULTS

A. What source?

We summarize the results of experiment 1 in Table II. Better detectors are obtained from the *Tagged Images* than from the *Tagged Videos*. When simply using all available tagged images for training, we obtain an mAP of 0.360, where for the video alternative we obtain an mAP of 0.219. We observe more genuine positives selected from the *Tagged Images* than the *Tagged Videos*. Selected training examples are shown in Figure 3. We attribute the difference in relative gain of tagged images over tagged video frames to their better annotation quality and to the fact that videos have more irrelevant content. When a tag is assigned to a complete video, often only a small fraction of the video content is relevant to the tag. We conclude from the results that tagged images are a better source than tagged videos for learning video concept detectors.

B. What strategy?

Figure 4 shows the performance curve of concept detectors derived from positive examples selected by the six strategies. Using images as training source is better than video for all strategies. Relevance selection is in general better then randomly selecting frames or images for all cut-offs in both sources as shown in Figure 4. For the best possible relevant image selection (fraction = 0.6), we gain a 7% relative improvement over the same amount of randomly selected images. For the video alternative, we obtain the best possible result for a fraction of 0.1, which obtains a 52% relative improvement over random selection. There is a clear decrease in performance as we select larger fractions of frames, see strategy 1a, Figure 4. Since larger fractions contain more noisy data, *i.e.* frames that also have low tag relevance score, the classifier learns less accurate models. In the case of tagged images there is less noise compared to tagged videos, since the tag is directly appointed to the visual content of the image. Consequently, for images larger fractions show better performance, see strategy 2a, Figure 4. Selection of smaller fractions also ignore relevant images, which results in training less accurate classifiers. It is also important to note that with a selection of relevant examples from the collections, apart from the gain in performance, we also reduce the number of training examples which leads to a speed-up during training.

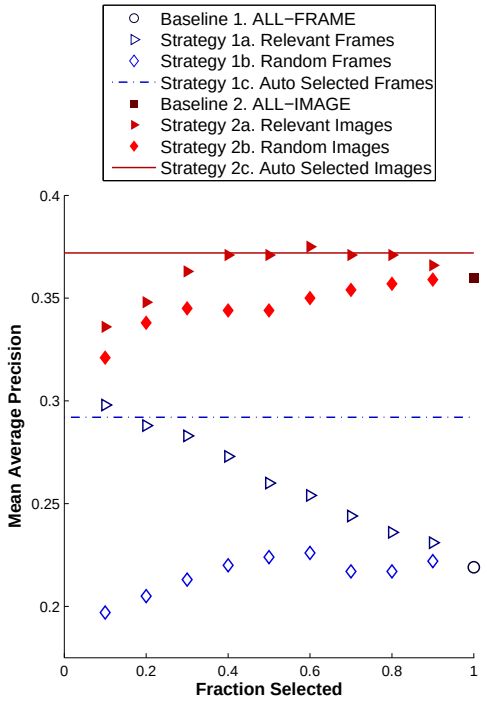

Fig. 4. **Experiment 2: What Strategy?** Training concept detectors with video frames or images, after tag relevance selection is always beneficial. The simple automatic selection is a good approximation of the best possible selected fraction for both video frames and images.

TABLE IV. EXPERIMENT 3: PRESENT-DAY COMPARISON.

	Strategies using tagged videos				Strategies using tagged images	
Concept	MIL [11]	MIL+ [4]	ALL-FRAME [16]	This Paper	ALL-IMAGE [12]	This Paper
Animal	0.032	0.040	0.053	0.081	0.088	**0.110**
Beach	0.053	0.079	0.278	0.276	0.386	**0.359**
Building	0.148	0.145	0.343	0.379	0.496	**0.531**
Car	0.130	0.130	0.164	0.239	0.249	**0.296**
Child	0.044	0.039	0.048	0.060	0.104	**0.111**
City	0.041	0.048	0.067	0.103	0.137	**0.158**
Face	0.357	0.376	0.359	0.504	**0.660**	0.651
Hand	0.081	0.076	0.098	0.102	0.169	**0.171**
Landscape	0.092	0.092	0.317	0.427	0.539	**0.539**
Mountain	0.030	0.030	0.079	0.193	0.495	**0.553**
Oceans	0.034	0.037	0.106	0.353	0.472	**0.490**
Outdoor	0.636	0.654	0.747	**0.824**	0.696	0.737
Plant	0.143	0.144	0.171	**0.263**	0.198	0.187
Road	0.147	0.147	0.193	0.242	**0.358**	0.336
Sky	0.229	0.242	0.311	0.423	**0.554**	0.516
Sports	0.041	0.040	0.119	0.145	0.127	**0.186**
Snow	0.028	0.032	0.092	0.217	0.289	**0.308**
Streets	0.103	0.102	0.119	0.150	0.202	**0.215**
Trees	0.372	0.407	0.490	0.548	**0.704**	0.681
Vehicle	0.128	0.138	0.225	0.313	0.282	**0.314**
mAP	0.175	0.180	0.219	0.298	0.360	**0.372**

The simple Bayes approach for automatic selection of positive examples, approximates the best selection quite closely (see the solid and dashed lines in Figure 4). In case of image selection it outperforms the best possible relevant image selection (fraction = 0.6) for 9 concepts even. We conclude that selecting relevant images and video frames is needed when learning concepts from the web. The selection cut off can be estimated automatically. We employed a simple Bayes approach, although more extensive sampling methods can be incorporated in the future.

C. Present-day comparison

In Table IV we compare our system with present-day approaches. When considering video as training source in a multiple instance learning setting, MIL+ [4] improves 3% over MIL [11]. The results confirm the value of context and temporal smoothing for identifying appropriate video tags at frame level. When we follow the ALL-FRAME strategy [16], which simply uses all frames of a tagged video, rather than selecting, we obtain even better results than the multiple instance approaches. The mAP for [16] results in a 22% relative improvement over [4]. However, the automatic selection for video frames outperforms all approaches for 19 out of 20 concepts. When relying on images as training source for concept detectors we observe similar behavior. The ALL-IMAGE strategy [12] reaches an mAP of 0.360, clearly improving over video and resulting in the best overall result for the concepts *Face*, *Road*, *Sky* and *Trees*. The automatic selection of example images is the best performer with an mAP of 0.372, and the top performer for 16 out of 20 concepts. We conclude that training concept detectors from a tagged image source using a relevant example selection strategy outperforms present-day alternatives.

VI. CONCLUSION

For learning video concepts from social media, *what visual tagging source is most suited for selecting positive training examples* is an important yet open question. This paper is, and to the best of our knowledge, the first endeavor to answer this question through a systematic empirical study. Additionally we also investigate *What strategy should be used for selecting positive examples from tagged sources?*, since it is known that social tags can be unreliable. Supported by experiments on a present day testbed, our major findings are: 1) Tagged images are the preferred choice as training source, when compared to tagged videos. Under all settings, concept detectors trained on tagged images surpass their counterparts trained on tagged videos, with an absolute improvement of approximately 10% in terms of mAP. While images may not be one's first option in the past research, we find that the better annotation quality let them beat videos with ease. 2) For both tagged videos and tagged images, selecting positive examples from those with the largest tag relevance scores is superior to getting positives at random. We show that relevant examples can be selected automatically, obtaining a near-optimal result. 3) Compared to several present day alternatives using all tagged examples, or Multiple Instance learning, the selection strategy of tag relevant examples produces better video concept detectors. To conclude, to learn video concept detectors from social media, we recommend relevance selection of tagged images.

VII. ACKNOWLEDGEMENT

This research is supported by the Dutch national program COMMIT and the Basic Research funds in Renmin University of China from the central government (13XNLF05).

REFERENCES

[1] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Enriching and localizing semantic tags in internet videos. In *MM*, 2011.

[2] S. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *MIR*, 2007.

[3] Y. Jiang, J. Yang, C. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *TMM*, 12(1):42–53, 2010.

[4] G. Li, M. Wang, Y.-T. Zheng, H. Li, Z.-J. Zha, and T.-S. Chua. Shottagger: tag location for internet videos. In *ICMR*, 2011.

[5] X. Li, C. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *TMM*, 11(7):1310–1322, 2009.

[6] X. Li, C. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *CIVR*, 2010.

[7] X. Li, C. Snoek, M. Worring, and A. W. M. Smeulders. Harvesting social images for bi-concept search. *TMM*, 14(4):1091–1104, 2012.

[8] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang. Tag ranking. In *WWW*, 2009.

[9] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[10] S. Ray and M. Craven. Supervised versus multiple instance learning: An empirical comparison. In *ICML*, 2005.

[11] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *NIPS*. 2008.

[12] A. Setz and C. Snoek. Can social tagged images aid concept-based video search? In *ICME*, 2009.

[13] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR*, 2006.

[14] Y. Sun and A. Kojima. A novel method for semantic video concept learning using web images. In *MM*, 2011.

[15] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *CIVR*, 2008.

[16] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. A system that learns to tag videos by watching youtube. In *ICVS*, 2008.

[17] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010.

[18] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtubecat: Learning to categorize wild web videos. In *CVPR*, 2010.

[19] J. Yang and A. Hauptmann. (un)reliability of video concept detection. In *CIVR*, 2008.

[20] S. Zhu, C.-W. Ngo, and Y.-G. Jiang. Sampling and ontologically pooling web images for visual concept learning. *TMM*, 14(4):1068–1078, 2012.