# Active Learning in Social Context for Image Classification

Elisavet Chatzilari[1,2],Spiros Nikolopoulos[1],Yiannis Kompatsiaris[1] and Josef Kittler[2]
[1]*Centre for Research & Technology Hellas - Information Technologies Institute*
[2]*Centre for Vision, Speech and Signal Processing, University of Surrey Guildford, UK*
{*ehatzi, nikolopo, ikom*}*@iti.gr, j.kittler@surrey.ac.uk*

Keywords: selective sampling, active learning, large scale, user-generated content, social context, image classification, multimodal fusion

Abstract: Motivated by the widespread adoption of social networks and the abundant availability of user-generated multimedia content, our purpose in this work is to investigate how the known principles of active learning for image classification fit in this newly developed context. The process of active learning can be fully automated in this social context by replacing the human oracle with the user tagged images obtained from social networks. However, the noisy nature of user-contributed tags adds further complexity to the problem of sample selection since, apart from their *informativeness*, our *confidence* about their actual content should be also maximized. The contribution of this work is on proposing a probabilistic approach for jointly maximizing the two aforementioned quantities with a view to automate the process of active learning. Experimental results show the superiority of the proposed method against various baselines and verify the assumption that significant performance improvement cannot be achieved unless we jointly consider the samples' *informativeness* and the oracle's *confidence*.

## 1 INTRODUCTION

The majority of state-of-the-art methods for automatic concept detection rely on the paradigm of pattern recognition through machine learning. Based on this paradigm, a model is parametrized to recognize all different attributes of a concepts' form and appearance using a set of training examples. The efficient estimation of model parameters mainly depends on two factors, the quality and the quantity of the training examples. High quality is usually accomplished through manual annotation, which is a laborious and time consuming task. This has a direct impact on the second factor since it inevitably leads into a small number of training examples and limits the performance of the generated models. In an effort to minimize the labelling effort, active learning (Cohn et al., 1994) trains the initial model with a very small set of labelled examples and enhances the training set by selectively sampling new examples from a much larger set of unlabelled examples (also referred as pool of candidates). These examples are selected based on their *informativeness*, i.e. how much they are expected to improve the model performance, and they are labelled by an oracle. They are typically found in the uncertainty areas of the model and their inclusion in the training set results in reducing the generalization error.

In the typical version of active learning, the pool of candidates usually consists of unlabelled examples that are annotated upon request by an errorless oracle. This requirement, which implies the involvement of a human annotator, renders active learning impractical in cases where the initial set needs to be enhanced with a significantly high number of additional samples while, at the same time, limits the scalability of this approach. On the other hand, the widespread use of Web 2.0 has made available large amounts of user tagged images that can be obtained at almost no cost and offer more information than their mere visual content. Our goal in this paper is to examine active learning in a rather different context from what has been considered so far. More specifically, if we could leverage these tags to become indicators of the images' actual content, we could potentially remove the need for a human annotator and automate the whole process. This, however, adds a new parameter, the oracle's *confidence* about the image's actual content, that should also be considered when actively selecting new samples. Additionally, even though in our case there is no annotation effort, adding informative instead of random samples is still important to mini-

mize the complexity of the classification models (i.e. achieve the same robustness with significantly fewer images).

The novelty of this work, in contrast to what has been considered so far in active learning, is to propose a sample selection strategy that maximizes not only the *informativeness* of the selected samples but also the oracle's *confidence* about their actual content. Towards this goal, we quantify the samples' *informativeness* by measuring their distance from the separating hyperplane of the visual model, while the oracle's *confidence* is measured based on the prediction of a textual classifier trained on a set of descriptors extracted using a typical bag of words approach (Joachims, 1998). Joint maximization is then accomplished by ranking the samples based on the probability to select a sample given the two aforementioned quantities (see Fig. 1). This probability indicates the benefit that our system is expected to have if the examined sample is selected to enhance the initial model. The rest of the manuscript is organized as follows. Section 2 reviews the related literature. In Section 3 the selective sampling algorithm is analysed and a theoretical analysis that quantifies the probability of selecting a new sample is presented. The experimental results are presented in Section 4 and conclusions are drawn in Section 5.

## 2 RELATED WORK

The examined context of this work combines three topics; active learning, multimedia domain and noisy data. During the past decade there have been many works exploring a subset of these topics, e.g. active learning in the multimedia domain (Wang and Hua, 2011), (Freytag et al., 2013) or active learning with noisy data (Settles, 2009), (Yan et al., 2011), (Fang and Zhu, 2012) or even non-active learning from noisy data in the multimedia domain (Chatzilari et al., 2012), (Raykar et al., 2010), (Yan et al., 2010), (Uricchio et al., 2013), (Verma and Jawahar, 2012), (Verma and Jawahar, 2013). However, it has been only recently that the scientific community started to investigate the implications of substituting the human oracle with a less expensive and less reliable source of annotations in the multimedia domain. There has been only a few attempts to combine active learning with user contributed images and most of them rely on either a human annotator or on the use of active crowdsourcing (i.e. a service like the MTurk) and not on passive crowdsourcing (i.e. the user provided tags that are typically found in social networks like flickr). In this direction, the authors of (Zhang et al., 2011) propose to use flickr notes in the typical active learning framework with the purpose of obtaining a training dataset for object localization. In a similar endeavour, the authors of (Vijayanarasimhan and Grauman, 2011) introduce the concept of *live learning* where they attempt to combine active learning with crowdsourced labelling. More specifically, rather than filling the pool of candidates with some canned dataset, the system itself gathers possibly relevant images via keyword search on flickr. Then, it repeatedly surveys the data to identify the samples that are most uncertain according to the current model, and generates tasks on MTurk to get the corresponding annotations.

On the other hand, social networks and user contributed content are leading most of the recent research efforts, mainly because of their ability to offer more information than the mere image visual content, coupled with the potential to grow almost unlimitedly. In this direction, the authors of (Li et al., 2013) propose a solution for sampling loosely-tagged images to enrich the negative training set of an object classifier. The presented approach is based on the assumption that the tags of such images can reliably determine if an image does not include a concept, thus making social sites a reliable pool of negative examples. The selected negative samples are further sampled by a two stage sampling strategy. First, a subset is randomly selected and then, the initial classifier is applied on the remaining negative samples. The examples that are most misclassified are considered as the most informative negatives and are finally selected to boost the classifier.

Our aim in this work is to investigate the extent to which the loosely tagged images that are found in social networks can be used as a reliable substitute of the human oracle in the context of active learning. Given that the oracle is not expected to reply with 100% correctness to the queries submitted by the selective sampling mechanism, we expect to face a number of implications that will question the effectiveness of active learning in noisy context. In this perspective our work differs from the large body of works that are found in the literature in the sense that most of them appear to be sensitive in label noise. In most of the works that do not use an expert as the oracle, MTurk is used instead to annotate the datasets. However, although active crowdsourcing services like MTurk are closer to expert's annotation (Nowak and Rüger, 2010) with respect to noise, they cannot be considered fully automated. In this work we rely on data originating from passive crowdsourcing (flickr images and tags) that although noisier, can be used to support a fully automatic active learning framework.

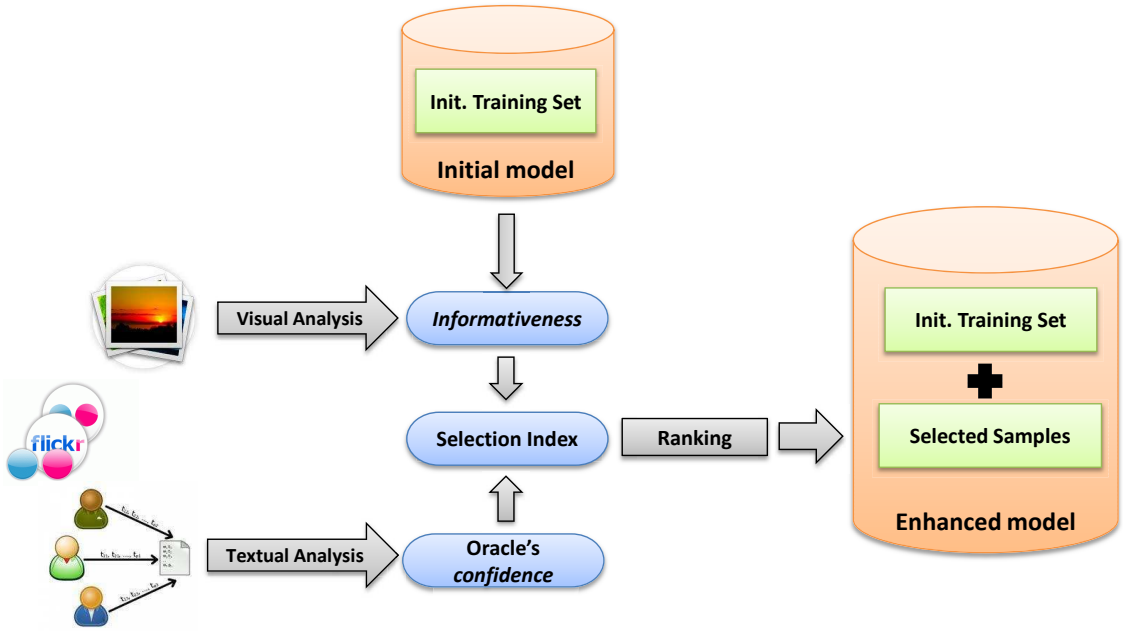


Figure 1: System Overview

The work presented in (Li et al., 2013) is examined under the same context as in this work (i.e. active learning in the multimedia domain using data from passive crowdsourcing), which, however, focuses on enriching the negative training set. Our work, on the other hand, focuses on enriching the positive training set that is more complex, since negative training samples are generally easier to harvest. Moreover, most of the existing datasets already contain a large number of negative examples but lack positives, which renders a positive sample selection strategy more applicable to a real world scenario.

# 3 SELECTIVE SAMPLING IN SOCIAL CONTEXT

Let us consider the typical case where, given a concept $c_k$, a base classifier is trained on the initial set of labelled images using Support Vector Machines (SVMs). We follow the popular rationale of SVM-based active learning methods ((Tong and Chang, 2001), (Campbell et al., 2000), (Schohn and Cohn, 2000)), which quantify the *informativeness* of a sample based on its distance from the separating hyperplane of the visual model (Section 3.1). In the typical active learning paradigm, a human oracle is employed to decide which of the selected informative samples are positive or negative. However, in the proposed scheme the human oracle is replaced with user contributed tags. Thus, in order to decide about a sample's actual label we utilize a typical bag-of-words classification scheme based on the image tags and the linguistic description of $c_k$. The outcome of this process is a confidence score for each image-concept pair (i.e. the oracle's *confidence*) which we consider as a strong indicator about the existence or not of $c_k$ in the image content (Section 3.2). Finally, the candidate samples are ranked based on the probability of selecting a new image given the two aforementioned quantities. The samples with the highest probability are considered the ones that jointly maximize the samples' *informativeness* and oracle's *confidence*, and are selected to enhance the initial training set.

## 3.1 Measuring informativeness

As already mentioned the *informativeness* of an image is measured using the distance of its visual representation from the hyperplane of the visual model. For the visual representation of the images, we have used the approach that was shown to perform best in (Chatfield et al., 2011). More specifically gray SIFT features were extracted at densely selected key-points at four scales, using the vl-feat library (Vedaldi and Fulkerson, 2008). Principal component analysis was applied on the SIFT features, decreasing their dimensionality from 128 to 80. The parameters of a Gaussian mixture model with $K = 256$ components were learned by expectation maximization from a set of descriptors, which were randomly selected from the en-

tire set of descriptors extracted by an independent set of images. The descriptors were encoded in a single feature vector using the Fisher vector encoding (Perronnin et al., 2010). Moreover, each image was divided in $1 \times 1, 3 \times 1, 2 \times 2$ regions, resulting in 8 total regions. A feature vector was extracted for each region by the Fisher vector encoding and the feature vector of the whole image ($1 \times 1$) was calculated using sum pooling (Chatfield et al., 2011). Finally the feature vectors of all 8 regions were $l2$ normalized and concatenated to a single $327680-dimensional$ feature vector, which was again power and $l2$ normalized.

For every concept $c_k$, a linear SVM classifier $(w_k, b_k)$, where $w_k$ is the normal vector to the hyperplane and $b_k$ the bias term, was trained using the labelled training set. The images labelled with $c_k$ were chosen as positive examples while all the rest were used as negative examples (One Versus All / OVA approach). For each candidate image $I_i$ represented by a feature vector $x_i$, the distance from the hyperplane $V(I_i, c_k)$ is extracted by applying the SVM classifier:

$$V(I_i, c_k) = w_k \times x_i^T + b_k \quad (1)$$

Using Eq. 1 we obtain the prediction scores, which indicate the certainty of the SVM model that the image $I_i$ depicts the concept $c_k$. In the typical self-training paradigm (Ng and Cardie, 2003), this certainty score is used to rank the samples in the pool of candidates and the samples with the highest certainty scores are chosen to enhance the models. However, as claimed and proven by the active learning theory (Settles, 2009), (Tong and Chang, 2001) these samples do not provide more information to the classifiers in order to alter significantly the classification boundaries.

Alternatively, as suggested by the active learning theory (Settles, 2009), the samples for which the initial classifier is more uncertain are more likely to increase the classifier's performance if selected. In the case of an SVM classifier, the margin around the hyperplane forms an uncertainty area and the samples that are closer to the hyperplane are considered to be the most informative ones (Fig. 2) (Tong and Chang, 2001). Based on the above, the samples that we want to select (i.e. the most informative) are the ones with the minimum distance to the hyperplane. Additionally, we only consider samples that lie in the margin area, since the rest of the samples are not expected to have any impact on the enhanced classifiers. We denote the probability to select an image $I_i$ given its distance to the hyperplane $V(I_i, c_k)$ as $P(S|V)$. Based on our previous observations, shown in Fig. 2, this probability can be formulated as a function of the sample's distance to the hyperplane which can be seen in Fig. 3:

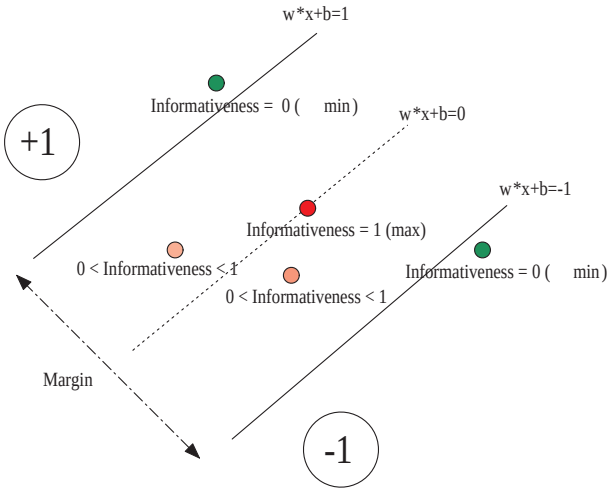


Figure 2: Informativeness

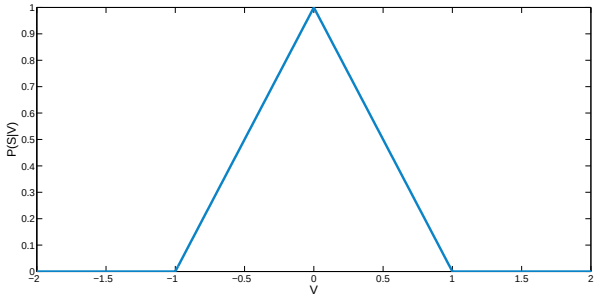


Figure 3: Probability of selecting a sample based on its distance to the hyperplane

$$P(S|V) = \begin{cases} 1 - |V| & \text{if } 0 < V < 1 \\ 0 & \text{else} \end{cases} \quad (2)$$

## 3.2 Measuring oracle's confidence

In order to measure the oracle's *confidence* about the existence of the concept $c_k$ in each tagged image, a typical bag-of-words scheme is utilized (Joachims, 1998). The vocabulary is extracted from a large independent image dataset crawled from flickr. Initially the distinct tags of all the images are gathered. The tags that are not included in WordNet are removed and the remaining tags compose the vocabulary. Then, in order to represent each image with a vector, a histogram is calculated by assigning the value 1 at the bins of the image tags in the vocabulary.

Afterwards, for every concept $c_k$ a linear SVM model $(w_k^{text}, b_k^{text})$ is trained using the tag histograms as the feature vectors. In order to do this, a training set of images that contains both tags and ground truth

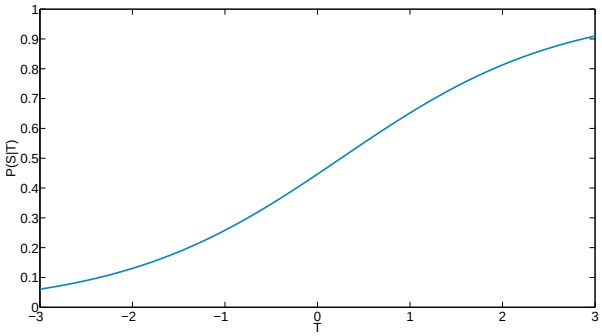


Figure 4: Probability of selecting a sample based on the oracle's confidence

information is utilized. The tags are required in order to calculate the feature vectors and the ground truth information to provide the class labels for training the model. In the testing procedure, for every tagged image $I_i$ the feature vector $f_i$ is calculated as above and the SVM model is applied. This results into a value for each tagged image $T(I_i, c_k)$, which corresponds to the distance of $f_i$ from the hyperplane:

$$T(I_i, c_k) = w_k^{text} \times f_i^T + b_k^{text} \quad (3)$$

This distance indicates the oracle's confidence that the examined image $I_i$ depicts the concept $c_k$.

We denote the probability to select an image $I_i$ given the oracle's confidence $T(I_i, c_k)$ as $P(S|T)$. In order to transform the oracle's confidence $T(I_i, c_k)$ (which corresponds to the distance of $I_i$ to the SVM hyperplane) into a probability we use a modification of Platt's algorithm (Platt, 1999) proposed by Lin et al. (Lin et al., 2007). Thus, the probability $P(S|T)$ can be formulated as a function of the oracle's confidence using the sigmoid function as shown in Fig. 4:

$$P(S|T) = \begin{cases} \frac{\exp(-AT-B)}{1+\exp(-AT-B)} & \text{if } AT+B \geq 0 \\ \frac{1}{1+\exp(AT+B)} & \text{if } AT+B < 0 \end{cases} \quad (4)$$

The parameters $A$ and $B$ are learned on the training set using cross validation.

### 3.3 Sample ranking and selection

Our aim is to calculate the probability $P(S=1|V,T)$, that an image is selected ($S=1$) given the distance of the image to the hyperplane $V$ and the oracle's confidence $T$. Considering that $V$ and $T$ originate from different modalities (i.e. visual and textual respectively) we regard them as independent. Using the basic rules of probabilities (e.g. Bayesian rule) and based on our assumption that V and T are independent we can express the probability $P(S|V,T)$ as follows:

$$\mathrm{P(S \mid V,T)} = \frac{\mathrm{P(V,T \mid S)\,P(S)}}{\mathrm{P(V,T)}} = \frac{\mathrm{P(S \mid V)}\frac{\mathrm{P(V)}}{\mathrm{P(S)}}\mathrm{P(S \mid T)}\frac{\mathrm{P(T)}}{\mathrm{P(S)}}\mathrm{P(S)}}{\mathrm{P(V,T)}} = \frac{\mathrm{P(S \mid V)\,P(S \mid T)\,P(V)\,P(T)}}{\mathrm{P(V,T)\,P(S)}}$$

In order to calculate the probability $P(S=1|V,T)$ and eliminate the probabilities $P(V)$, $P(T)$ and $P(V,T)$, we divide the probability of selecting an image with the probability of not selecting it.

$$\frac{P(S=1|V,T)}{P(S=0|V,T)} = \frac{\frac{P(S=1|V)P(S=1|T)P(V)P(T)}{P(V,T)P(S=1)}}{\frac{P(S=0|V)P(S=0|T)P(V)P(T)}{P(V,T)P(S=0)}} \Rightarrow$$

$$\frac{P(S=1|V,T)}{P(S=0|V,T)} = \frac{\frac{P(S=1|V)P(S=1|T)}{P(S=1)}}{\frac{P(S=0|V)P(S=0|T)}{P(S=0)}}$$

Then we use the basic probabilistic rule that the probability of an event's complement equals 1 minus the probability of the event ($P(S=0|V,T) = 1 - P(S=1|V,T)$).

$$\frac{\mathrm{P(S=1 \mid V,T)}}{1 - \mathrm{P(S=1 \mid V,T)}} = \frac{\frac{\mathrm{P(S=1|V)P(S=1|T)}}{\mathrm{P(S=1)}}}{\frac{(1-\mathrm{P(S=1|V)})(1-\mathrm{P(S=1|T)})}{1-\mathrm{P(S=1)}}} \Rightarrow \ldots \Rightarrow$$

$$P(S=1|V,T) = \frac{P(S=1|V)P(S=1|T)}{P(S=1) - P(S=1)P(S=1|T)} \cdots$$
$$\frac{(1-P(S=1))}{-P(S=1)P(S=1|V) + P(S=1|V)P(S=1|T)} \quad (5)$$

Thus we only need to estimate three probabilities: $P(S=1)$, $P(S=1 \mid V)$ and $P(S=1 \mid T)$. The first one is set to 0.5 as the probability of selecting an image without any prior knowledge is the same with the probability of dismissing it. For the estimation of the other two probabilities we use the equations 2 and 4 (shown in Fig. 3 and 4). Finally, the top $N$ images with the highest probability $P(S=1|V,T)$ are selected to enhance the initial training set.

## 4 EXPERIMENTS

### 4.1 Datasets and implementation details

Two datasets were employed for the purpose of our experiments. The imageCLEF dataset *IC* (Thomee

and Popescu, 2012) consists of 25000 labelled images and was split into two parts (15k train and 10k test images). The ground truth labels were gathered using Amazon's crowdsourcing service MTurk. The dataset was annotated by a vocabulary of 94 concepts which belong to 19 general categories (*age, celestial, combustion, fauna, flora, gender, lighting, quality, quantity, relation, scape, sentiment, setting, style, time of day, transport, view, water, weather*). On average there are 934 positive images per concept, while the minimum and the maximum number of positive images for a single concept is 16 and 10335 respectively. In our experimental study the 15k training images were used to train the initial classifiers.

The MIRFLICKR-1M dataset $F$ (Mark J. Huiskes and Lew, 2010) consists of one million loosely tagged images harvested from flickr. The images of $F$ were tagged with 862115 distinct tags of which 46937 were meaningful (included in WordNet). After the textual preprocessing, i.e. removing the tags that were not included in WordNet, 131302 images had no meaningful tags, 825365 images were described by 1 to 16 meaningful tags and 43333 images had more than 16 meaningful tags. Given that the *IC* dataset is a subset of $F$, the images that are included in both sets were removed from $F$. In our experiments, this dataset constitutes the pool of loosely tagged images, out of which the top $N = 500$ images ranked by Eq. 5 are selected for each concept (i.e. 94 concepts * 500 images per concept = 47k images total) to act as the positive examples enhancing the initial training set. Finally, mean average precision (MAP) served as the metric for measuring the models' classification performance and evaluating the proposed approach.

## 4.2 Evaluation of the proposed selective sampling approach

The objective of this section is to compare the proposed active sample selection strategy against various baselines. The first baseline is the initial models that were generated using only the ground truth images from the training set (15k images). Afterwards, the initial models are enhanced with positive samples from $F$ using the following sample selection strategies:

**Self-training (Ng and Cardie, 2003)** The images that maximize the certainty of the SVM model trained on visual information (i.e. maximize the visual distance to the hyperplane as measured by Eq. 1) are chosen.

**Textual based** The images that maximize the oracle's *confidence* are selected (Eq. 4).

**Max *informativeness*** The images that maximize the *informativeness* (i.e. are closer to the hyperplane) are chosen (Eq. 2).

**Naïve oracle** The images that maximize the *informativeness* (Eq. 2) and explicitly contain the concept of interest in their tag list are chosen (i.e. plain string matching is used).

**Proposed approach** The images that jointly maximize the sample's *informativeness* and the oracle's *confidence* are chosen (Eq. 5).

The average performance of the enhanced classifiers using the aforementioned sample selection strategies is shown in Table 1. We can see that in all cases the enhanced classifiers outperform the baseline. Moreover, the approaches relying on active learning yield a higher performance gain compared to the typical self-training approach, showing that the *informativeness* of the selected samples is a critical factor. The same conclusion is drawn when comparing the textual based approach to the proposed method, showing that *informativeness* is crucial to optimize the learning curve, i.e. achieve higher improvement when adding the same number of images. On the other hand, the fact that the proposed sample selection strategy and the string matching variation (i.e. naïve oracle) outperform significantly the visual-based variations, verifies that the oracle's *confidence* is a critical factor when applying active learning in social context and unless we manage to consider this value jointly with *informativeness*, the selected samples are inappropriate for improving the performance of the initial classifiers.

Additionally, we note that the naïve oracle variation performs relatively well, which can be attributed to the high prediction accuracy achieved by string matching. Nevertheless, the recall of string matching is expected to be lower than the textual similarity algorithm used in the proposed approach (Section 3.2), since it does not account for synonyms, plural versions and the context of the tags. This explains the superiority of our method compared to the naïve oracle variation. In order to verify that the performance improvement of the proposed approach compared to the naïve oracle is statistically significant, we apply the Student's t-test to the results, as it was proposed for significance testing in the information retrieval field (Smucker et al., 2007). The obtained p-value is 2.58e-5, significantly smaller than 0.05, which is typically the limit for rejecting the null hypothesis (i.e. the results are obtained from the same distribution and thus the improvement is random), in favour of the alternative hypothesis (i.e. that the obtained improvement is statistically significant).

Table 1: Performance scores

| Model | mAP (%) |
|---|---|
| Baseline | 28.06 |
| Self-training | 28.68 |
| Textual based | 29.89 |
| Max *informativeness* | 28.73 |
| Naïve oracle | 30 |
| **Proposed approach** | **31.22** |

Moreover, a per concept comparison of the enhanced models generated by the two best performing approaches of Table 1 (i.e. the proposed approach and the naïve oracle variation) to the baseline classifiers can be seen in the bar diagram shown in Fig. 5. We can see that the proposed approach outperforms the naïve oracle in 70 concepts out of 94. It is also interesting to note that the naïve oracle outperforms the proposed approach mostly in concepts that depict objects such us amphibian-reptile, rodent, baby, coast, cycle and rail. This can be attributed to the fact that web users tend to use the same keywords to tag images with concepts depicting strong visual content, which are typically the object of interest in an image. In such cases, the string matching oracle can be rather accurate, providing valid samples for enhancing the classifiers. On the other hand, the proposed approach copes better with more abstract and ambiguous concepts for which the context is a crucial factor (e.g. flames, smoke, lens effect, small group, co-workers, strangers, circular wrap and overlay).

A closer look at the obtained results from the proposed approach shows that the concept with the most notable increase in performance is the *spider*, initially trained by 16 positive examples yielding only 5.48% AP. After adding the samples that were indicated by the proposed oracle, the classifier gains 23.31 units of performance, resulting in 28.79% average precision. Similarly, other concepts yielding a performance gain in the range of 5 and more units include *stars, rainbow, flames, fireworks, underwater, horse, insect, baby, rail* and *air*. Most of these concepts' baseline classifiers yield a low performance. Another category of concepts are the ones with slight variations on performance, below 0.1%. This category includes the concepts *cloudy sky, coast, city, tree, none, adult, female, no blur* and *city life* whose baseline classifiers yield a rather high performance and are trained with 3600 positive images on average. This shows that the proposed method, as it could be expected, is more beneficial for difficult concepts, i.e. whose initial classifiers perform poorly. Finally, there are also the concepts that either yield minor variations or even decrease in performance and consist in *melancholic, unpleasant* and *big group*. This can be attributed to the ambiguous nature of these concepts which renders the oracle unable to effectively determine their existence.

### 4.3 Comparing with state-of-the-art

In this section the proposed approach is compared to the methods submitted to the 2012 ImageClef competition (Thomee and Popescu, 2012) and specifically in the concept annotation task for *visual concept detection, annotation, and retrieval using Flickr photos*[1]. Since the proposed approach is only using the visual information of the test images without taking into account the associated tags, it is only compared to the visual-based approaches submitted in the competition. The performance scores for the three metrics utilized by the competition organizers (miAP, GmiAP and F-ex) are reported in Table 2 for each of the 14 participating teams, along with the baselines of Table 1 and the proposed approach. In order to measure the F-ex score, the threshold for the positive-negative class separation was set to zero, i.e. images with an SVM prediction score greater than zero were annotated as positive and negative otherwise. We can see that our approach is ranked third in terms of miAP, first in terms of GmiAP and fifth in terms of F-ex. Additionally, we note that the proposed approach outperforms the rest in terms of GmiAP, which according to (Thomee and Popescu, 2012) is a metric susceptible to better performances on difficult concepts. This explains the superiority of our approach and the higher performance gain compared to our baseline since it tends to improve the performance of the difficult concepts, as it was also observed in Section 4.2 (see Fig. 5). Moreover, it is important to note that the proposed approach has achieved these very competitive scores by using a single feature space (gray SIFT features), which was not the case for the other participants that relied on more than one feature spaces (Thomee and Popescu, 2012).

## 5 CONCLUSIONS

In this paper, we propose an automatic variation of active learning for image classification adjusted in the context of social media. This adjustment consists in replacing the typical human oracle with user tagged images obtained from social sites and in using a probabilistic approach for jointly maximizing the *informativeness* of the samples and the oracle's *confidence*. The results show that in this context it

[1] http://imageclef.org/2012/photo-flickr

Table 2: Comparison with ImageClef 2012

| Team | miAP | GmiAP | F-ex |
|---|---|---|---|
| LIRIS | **34.81%** | **28.58%** | 54.37% |
| NPDILIP6 | 34.37% | 28.15% | 41.99% |
| NII | 33.18% | 27.03% | **55.49%** |
| ISI | 32.43% | 25.90% | 54.51% |
| MLKD | 31.85% | 25.67% | 55.34% |
| CERTH | 26.28% | 19.04% | 48.38% |
| UAIC | 23.59% | 16.85% | 43.59% |
| BUAA AUDR | 14.23% | 8.18% | 21.67% |
| UNED | 10.20% | 5.12% | 10.81% |
| DBRIS | 9.76% | 4.76% | 10.06% |
| PRA | 9.00% | 4.37% | 25.29% |
| MSATL | 8.68% | 4.14% | 10.69% |
| IMU | 8.19% | 3.87% | 4.29% |
| URJCyUNED | 6.22% | 2.54% | 19.84% |
| Baseline | 30.37% | 24.21% | 48.6% |
| Self-training | 30.77% | 24.41% | 49.63% |
| Textual based | 32.48% | 26.84% | 51.7% |
| Max *informativeness* | 30.83% | 24.48% | 52.24% |
| Naïve oracle | 32.18% | 26.53% | 51.66% |
| **Proposed approach** | **33.84%** | **29.17%** | **52.64%** |

is critical to jointly consider these two quantities for successfully selecting additional samples to enhance the initial training set. Additionally, we noticed that the naïve oracle performs very well on concepts that depict strong visual content corresponding to typical foreground visual objects (e.g fish, spider, bird and baby), while the proposed approach copes better with more abstract and ambiguous concepts (e.g. flames, smoke, strangers and circular wrap), since the utilized textual classifier accounts for the context of the tags as well.

Finally, an interesting note is that the difficult concepts (i.e. models with low performance) tend to gain much more in terms of effectiveness from such bootstrapping methods, as shown in Fig. 5. Similar conclusions are drawn when comparing the proposed approach, which trained a simple SVM classifier using a single feature space to the more sophisticated approaches of the ImageCLEF 2012 challenge, which typically used many feature spaces. Especially in the case of difficult concepts, as shown by the superiority of the proposed approach based on the GmiAP metric, we can also conclude that it is more important to find more positive samples than more sophisticated algorithms.

Our plans for future work include the use of flickr groups as a richer and more large-scale pool of candidates for positive samples and the extension of the proposed approach to an on-line continuous learning scheme.

## Acknowledgements

This work was supported by the EU 7th Framework Programme under grant number IST-FP7-288815 in project Live+Gov (www.liveandgov.eu).

## REFERENCES

Campbell, C., Cristianini, N., and Smola, A. J. (2000). Query learning with large margin classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 111–118, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*.

Chatzilari, E., Nikolopoulos, S., Kompatsiaris, Y., and Kittler, J. (2012). Multi-modal region selection approach for training object detectors. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 5:1–5:8, New York, NY, USA. ACM.

Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. *Mach. Learn.*, 15(2):201–221.

Fang, M. and Zhu, X. (2012). I don't know the label: Active learning with blind knowledge. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2238–2241.

Freytag, A., Rodner, E., Bodesheim, P., and Denzler, J. (2013). Labeling examples that matter: Relevance-based active learning with gaussian processes. In Weickert, J., Hein, M., and Schiele, B., editors, *GCPR*, volume 8142 of *Lecture Notes in Computer Science*, pages 282–291. Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Ndellec, C. and Rouveirol, C., editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg.

Li, X., Snoek, C. G. M., Worring, M., Koelma, D. C., and Smeulders, A. W. M. (2013). Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia*, In press.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276.

Mark J. Huiskes, B. T. and Lew, M. S. (2010). New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, pages 527–536, New York, NY, USA. ACM.

Ng, V. and Cardie, C. (2003). Bootstrapping coreference classifiers with multiple machine learning algorithms.

In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 113–120.

Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA. ACM.

Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 143–156. Springer-Verlag.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322.

Schohn, G. and Cohn, D. (2000). Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 839–846, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Smucker, M. D., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 623–632.

Thomee, B. and Popescu, A. (2012). Overview of the clef 2012 flickr photo annotation and retrieval task. in the working notes for the clef 2012 labs and workshop. Rome, Italy.

Tong, S. and Chang, E. (2001). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA '01, pages 107–118, New York, NY, USA. ACM.

Uricchio, T., Ballan, L., Bertini, M., and Del Bimbo, A. (2013). An evaluation of nearest-neighbor methods for tag refinement.

Vedaldi, A. and Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.

Verma, Y. and Jawahar, C. V. (2012). Image annotation using metric learning in semantic neighbourhoods. In *Proceedings of the 12th European conference on Computer Vision - Volume Part III*, ECCV'12, pages 836–849.

Verma, Y. and Jawahar, C. V. (2013). Exploring svm for image annotation in presence of confusing labels. In *Proceedings of the 24th British Machine Vision Conference*, BMVC'13.

Vijayanarasimhan, S. and Grauman, K. (2011). Large-scale live active learning: Training object detectors with crawled data and crowds. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1449 –1456.

Wang, M. and Hua, X.-S. (2011). Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol.*, 2(2):10:1–10:21.

Yan, Y., Rosales, R., Fung, G., and Dy, J. (2011). Active learning from crowds. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1161–1168, New York, NY, USA. ACM.

Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., and Dy, J. (2010). Modeling annotator expertise: Learning when everybody knows a bit of something.

Zhang, L., Ma, J., Cui, C., and Li, P. (2011). Active learning through notes data in flickr: an effortless training data acquisition approach for object localization. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 46:1–46:8, New York, NY, USA. ACM.

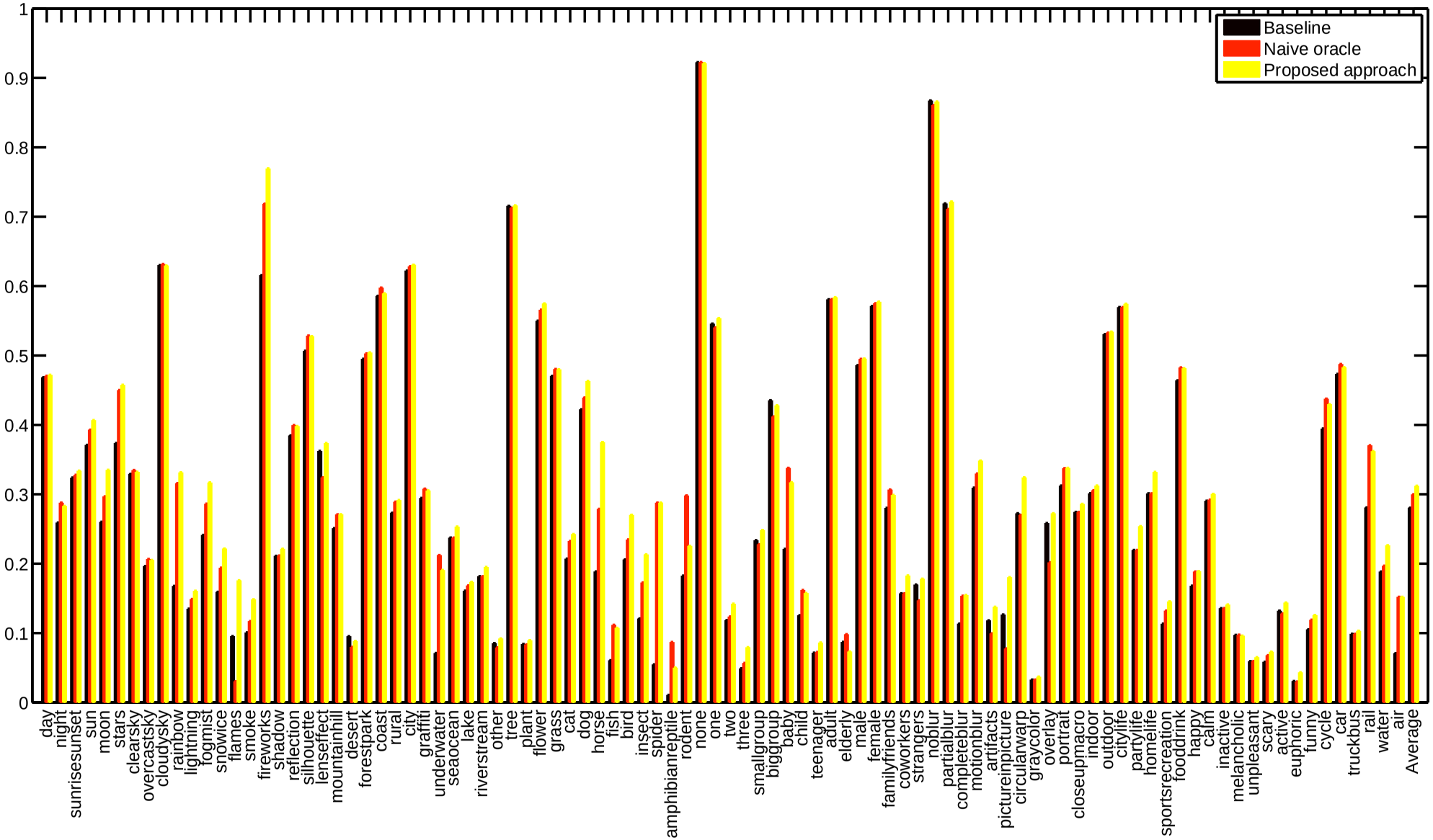


Figure 5: Per concept comparison of the two best performing approaches (i.e. the naïve oracle and the proposed approach) to the baseline (best viewed in colour)