Learning Visual Contexts for Image Annotation From Flickr Groups

Adrian Ulges, Marcel Worring, *Member, IEEE*, and Thomas Breuel, *Member, IEEE*

***Abstract*—We present an extension of automatic image annotation that takes the *context* of a picture into account. Our core assumption is that users do not only provide individual images to be tagged, but group their pictures into batches (e.g., all snapshots taken over the same holiday trip), whereas the images within a batch are likely to have a common style. These batches are matched with categories learned from Flickr groups, and an accurate context-specific annotation is performed.**

In quantitative experiments, we demonstrate that Flickr groups, with their user-driven categorization and their rich group space, provide an excellent basis for learning context categories. Our approach—which can be integrated with virtually any annotation model—is demonstrated to give significant improvements of above 100%, compared to standard annotations of individual images.

***Index Terms*—Content-based image retrieval, context, image annotation.**

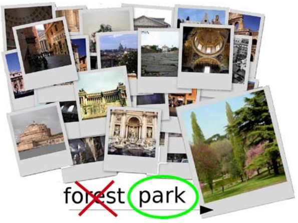


Fig. 1. Annotating a batch of pictures from a holiday trip to Rome: while annotating the bottom right image with a global "world" model is difficult (potential tags might be "forest" or "park"), a context-specific model trained on the Flickr group "Rome Pictures" can identify the correct tag to be "park".

I. Introduction

IMAGE annotation is a core challenge in image retrieval and has been subject to intensive study over the last decade [1], [5], [13], [19], [22]. It is concerned with automatically enriching images with textual annotations (or *tags*) describing objects, locations, and actions appearing in them. If we were able to reliably provide descriptive labels from a large-scale vocabulary, we could improve the ease and speed with which we access and manage images significantly. However, image annotation poses an extraordinarily difficult challenge due to large tag vocabularies encountered, enormous intra-class variation, and a lack of large-scale high-quality training data. Correspondingly, current systems suffer from two key problems: a *scalability* problem (in particular, we require larger-scale training data and robust strategies to learn thousands of tags), and an *accuracy* problem requiring quality to be driven closer towards the level of human annotations.

To face the challenge of scalability, a prominent trend over the last years has been to employ web-based image collections: services like Flickr and web search engines allow us to collect huge amounts of diverse content, which is often enriched with user-generated tags and meta-data. By employing this information for training, annotation systems have been shown to perform autonomous visual learning at a large scale [24], [36], [44], [45].

In this paper, we will pursue the idea of web-based learning further and demonstrate that information we find online can not only be used to face the *scalability* problem but also the *accuracy* problem. To do so, we will exploit the fact that users *structure* the content they share on the web: for example, the Flickr community collaboratively organizes its pictures in a dynamic, user-driven collection of categories called *Flickr Groups*. Currently, over 200 000 such groups have been defined, related to all kinds of topics like "sightseeing trips", "party pictures", or "nature photography".

While previous work has addressed a *textual* learning from this group information [34], [35], it has not been studied from a computer vision perspective so far. In this paper, we will fill this gap and present a way to improve the accuracy of image annotation by employing Flickr group information.

Our key idea is that—instead of training a single global annotation model on all images we download from Flickr [24], [36], [44], [45]—we can obtain much more accurate annotations by training *specific* models on distinct Flickr groups. This idea is illustrated in Fig. 1: an ambiguous picture is shown for which a global annotation model might assign either the tag "forest" or "park". However, when adding the information that the picture belongs to a batch of images taken over a trip to Rome, we can use a "Rome-specific" annotation model, which represents the appearance and frequency of tags in the batch better and thus allows us to infer the correct tag "park".

This example illustrates that disambiguation can be achieved by taking the context of a picture into account. In our case, the term "context" refers to other images in the same batch,

Manuscript received June 10, 2010; revised October 25, 2010; accepted December 07, 2010. Date of publication December 20, 2010; date of current version March 18, 2011. This work was supported by the German Research Foundation (DFG) under project MOONVID (BR 2517/1-1). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Svetha Venkatesh.

A. Ulges is with the German Research Center for Artificial Intelligence, D-67663 Kaiserslautern, Germany (e-mail: adrian.ulges@dfki.de).

M. Worring is with the Intelligent Systems Lab, University of Amsterdam, 1098 GH Amsterdam, The Netherlands (e-mail: m.worring@uva.nl).

T. Breuel is with the Department of Computer Science, University of Kaiserslautern, D-67663 Kaiserslautern, Germany (e-mail: tmb@cs.uni-kl.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2101051

for example taken over the same holiday trip. Context information is frequently available in practice, as users organize and group their pictures in folders or web albums. We focus on improving image annotation by exploiting this information: during learning, we will rely on Flickr group information for building group-specific annotation models, and during annotation, group information is assumed to be provided by the user (alternatively, it can be inferred from capture time or location [4]). Our key hypothesis is that—*if* such information is given—it improves image annotation distinctly.

In the following, we will study this approach of matching image contexts with Flickr groups, and validate significant relative accuracy improvements of more than 100% over an annotation of individual images that constitutes the state of the art. We will also demonstrate that Flickr groups provide an excellent basis for learning visual contexts: not only is this information freely available and driven by a large community of users, but it is also extremely rich, covering a wide range of fine-grain semantic categories. These aspects will be shown to be essential for a successful context learning.

Finally, it should also be noted that our work does not present a new statistical annotation model. Rather, our approach can easily be used as an extension to existing approaches like [1], [5], [13], [19], and [22] such that context information leads to improved accuracy.

This paper is organized as follows: we will first discuss related work in Section II. After this, we introduce the proposed approach in detail (Section III), and demonstrate for two standard annotation models (namely, a simple texton model and the "PLSA-Words" model [30]) how an annotation of individual images can be extended to a context-based one by learning from Flickr groups. In Section IV, we present quantitative experiments validating strong improvements by the proposed approach. Section VI gives a conclusion.

II. Related Work

This section outlines research contributions related to our approach. We will first introduce the field of image annotation in general (Section II-A). After this, specific approaches close to the presented work are described in more detail: methods exploiting web content for an automated training (Section II-B) and related strategies for using context in image annotation (Section II-C).

A. Image Annotation

Image annotation has been studied intensively over the last decade, and a variety of approaches has been presented for assigning tags to images [5], [13], [22], [30] or even to specific regions inside them [9], [19], [31]. Usually, images are viewed as collections of local regions, which can be obtained using an image segmentation approach [21] or by a sampling of local patches [13]. The goal of image annotation is to map this description to tags from a pre-defined vocabulary.

Different approaches have been taken towards this challenge: some popular ones are *generative models*, which are targeted at estimating a joint distribution of local image features and annotations. This can be achieved using probability tables [12], [31], *topic models* [1], [30], or *relevance models* [13], [20], [21]. Further, Carneiro *et al.* [5] proposed a continuous model based on Gaussian mixtures.

Other approaches have employed nearest neighbor matching techniques [22], [24], [36], [45]—i.e., similar training images are found and their tags are transferred—or used discriminative classifiers like maximum entropy [18] or support vector machines (SVMs) [28].

Alternatively, tags can be assigned to specific regions in the image. In this case, the fact that in training, usually no explicit correspondences between tags and image regions are provided renders image annotation a *weakly supervised learning* problem. Duygulu *et al.* [9] drew an analogy to latent correspondences in machine translation and adopted the EM algorithm for tag-region alignment. Yang *et al.* applied multiple-instance learning [46], and Kück *et al.* [19] formulated a constrained semi-supervised learning problem in a probabilistic framework.

While all these methods differ in terms of features and underlying statistical models, a common limitation is that images are treated as independent samples. This assumption is frequently violated in practice, as personal photos tend to come as series of coherent pictures. In our method, we employ this context to improve annotation. As our approach is general, it can complement standard image models as the ones introduced above.

B. Exploiting Web-Based Information

A key challenge for image annotation is the acquisition of proper training data for large-scale tag vocabularies. For this purpose, researchers have turned towards the web, where useful information in form of text, images, and video is freely available.

Web-based text has been exploited to learn relations between different terms using tag co-occurrence statistics. This information can be exploited for post-processing annotation results, and has been acquired, for example, from Flickr [47].

Image information from the web has been exploited for training annotations systems as well, whereas labeled pictures are downloaded from photo sharing websites like Flickr [24], [25], [36], [38] or via web image search engines [14], [23]. This information is often combined with a nearest neighbor tagging, i.e., Flickr images similar to a target picture are found and their tags are transferred [24], [25], [36]. Finally, for the video domain, video sharing portals like YouTube have been investigated as training data similar to the use of web image content [17], [42].

What distinguishes our work from these previous efforts is the use of web-based *category* information: we do not only employ Flickr content as training data but also exploit the fact that users sort their pictures into Flickr groups. This information source has been studied before from a textual perspective by Negoescu *et al.* [34], [35], who analyzed tags to cluster groups to "hypergroups" and thus achieve a better understanding of the Flickr group space. From an image annotation perspective, previous work has exploited Flickr *user* information for refining noisy training data [26], but to the best of our knowledge, our work is the first to exploit Flickr group information for context-based annotation.

C. Image Annotation Using Context

While the majority of current image annotation approaches treats pictures individually (see Section II-A), some previous work has exploited information conveyed in groups of images and thus made use of context. Often, context is associated with certain *events* over which pictures have been taken. Gallagher *et al.* [15] match groups of images with events in personal calendars (like "George's Wedding"). Naaman *et al.* [33] improve person identification by grouping pictures to events using meta-data like capture time and location. In a similar fashion, Cao *et al.* [4] cluster pictures and perform an event-based image annotation, where context is used in form of correlation terms between events and image tags. Finally, Cristani *et al.* [7] addressed the geo-recognition problem (i.e., the inference of the geographic location where a picture was taken) and demonstrated improvements by using a whole group of images instead of individual ones.

What we learn from these approaches is that pictures can be aggregated to groups using meta-data such as capture time and location. We will rely on this very same grouping information and focus on how to exploit it for annotation purposes. However, what distinguishes our work is that previous contributions have not provided a rigorous way for *learning* contexts at a large scale (training information for different categories of contexts has been acquired manually, which is infeasible in many practical situations). To overcome this problem, we exploit Flickr groups for an automatic and scalable learning of contexts.

III. Approach

This section introduces the proposed visual learning from Flickr groups in detail. The general approach is illustrated in Fig. 2: training information is downloaded from Flickr, and specific annotation models are trained for different Flickr groups (like "Rome" or "Wedding"). The user provides a new batch of images to be annotated. For this batch, the most appropriate Flickr group is chosen, and the group-specific model is used for an accurate annotation. In the following, we will first motivate our choice of Flickr groups as an information source, pointing out several key characteristics that make them suitable for learning context categories (Section III-A). Afterwards, our context-based approach will be introduced as an extension to image annotation that can be integrated with a variety of base models (Section III-B), and two such extensions of well-known image annotation models are presented, respectively, for the texton model, Section III-C, and a PLSA-based model, Section III-D.

A. Why Flickr Groups?

Our general idea is to learn categories of contexts and then match pictures to be annotated with these categories. It is a key question *which data source* to employ for such context learning, and we suggest Flickr groups as an answer. In the following, we motivate this choice, pointing out several key characteristics of Flickr groups that make them suitable for an effective learning of meaningful visual contexts.

- **Availability**: First—and most obviously—Flickr group information is freely available, such that a fully automatic learning can be performed.

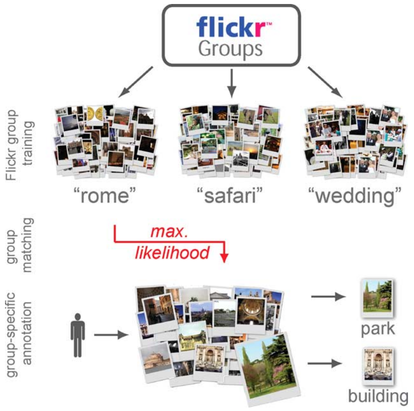


Fig. 2. Illustration of the proposed visual learning from Flickr groups: different annotation models are trained for different Flickr groups. These models are matched with an image batch provided by the user. Finally, this batch is annotated using a group-specific model.

- **User-Driven Categorization**: Grouping pictures into semantic categories is a difficult challenge, as semantically related content may be highly diverse (think of the category "Hiking" showing panoramas of diverse landscapes as well as group pictures with friends). As humans are good at making these connections, it is a positive aspect that Flickr groups are structured by a large community of human users.
- **Richness of Group Space**: In practice, a users' target pictures may be very specific and differ from any context category we have learned previously. Still, we would like to find an *appropriate* (if not perfect) model for a user's pictures. We expect that the more contexts we learn (i.e., the *richer* our space of categories is), the better we find a match to explain a specific batch of pictures. Here, Flickr—with its extremely rich collection of more than 200 000 groups [34]—is well suited for a good generalization of context learning.

B. Basic Concepts

For representing pictures' visual content, we use the well-known "bag-of-visual-words" approach [39]: each image is viewed as a collection of local features, which are quantized into clusters $v \in \{1, \dots, V\}$ called "visual words". An image I is then represented by a histogram $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_V)$, where $\mathbf{x}_v$ denotes the number of occurrences of visual word v in I.

Let us now move to the annotation step, where a group of test pictures $\mathcal{X}'$ is to be labeled. Again, each image I' is represented by a bag-of-visual-words histogram $\mathbf{x}'$. The goal of image annotation is to infer probabilities (or *scores*) $P(t|I')$ indicating for each $I' \in \mathcal{X}'$ whether tags t from a pre-defined vocabulary T are adequate annotations.

This process is based on a statistical model ϕ learned in a previous training step. We assume a set of training images $\mathcal{X}$ to be given, which are downloaded from Flickr. Each image $I \in \mathcal{X}$ comes with a set of annotations (or *tags*) $\mathbf{t} \subset T$ (which Flickr users have provided with the picture). Our key novelty is that we further assume training images to be divided into categories $g \in G$, i.e., $\mathcal{X} = \cup_{g \in G} \mathcal{X}^g$. In practice, these categories correspond to different Flickr groups—we have a group $g_1 =$

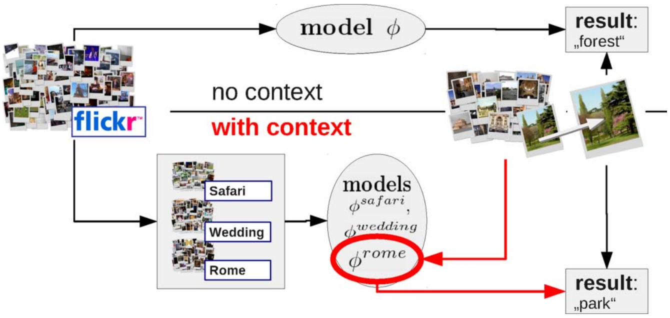


Fig. 3. Comparing standard image annotation ("no context", top) with the proposed context-based extension ("with context", bottom): the standard approach derives a single global annotation model ϕ from all training images, and treats images from the test batch as independent samples. In contrast to this, the proposed approach learns specific models ϕ^{rome}, ϕ^{safari}, etc., and uses the context of a picture to pick an adequate specific model. This use of context is independent of the concrete type of annotation model ϕ and can thus be integrated with a variety of approaches.

"wedding pictures", another one g_2 = "pictures of Rome", etc.

Let us now discuss how to make use of this grouping information (see Fig. 3 for an illustration). Standard image annotation ("no context")—which does not take group information into account—would derive a *global* model ϕ from *all* training images $\mathcal{X}$ and then use ϕ to label all pictures $I' \in \mathcal{X}'$ individually: $P(t|I') \approx \phi(\mathbf{x}', t)$. In contrast to this, the proposed approach proceeds in three steps (as illustrated in Fig. 2).

1) **Flickr Group Training**: We learn separate models ϕ^g, one for each group $\mathcal{X}^g$.
2) **Group Matching**: Given a batch of test images $\mathcal{X}'$, we decide which Flickr group fits the visual appearance of $\mathcal{X}'$ best. This is done using a maximum-likelihood criterion:

$$g^* = \arg\max_{g \in G} \; p(\mathcal{X}'|\phi^g). \tag{1}$$

3) **Group-Specific Annotation**: The group-specific model ϕ^{g^*} is used to infer tag scores for each image $I' \in \mathcal{X}'$, i.e., $P(t|I') \approx \phi^{g^*}(\mathbf{x}', t)$. This is expected to give much more accurate results than using a global model ϕ.

This approach can be seen as an adaptation of the *style consistency model*, which has been introduced by Sarkar and Nagy in the domain of optical character recognition [37]. It is based on two key assumptions: first, we expect that the accuracy of group matching increases when using context. While matching a single picture to a Flickr group is highly ambiguous, a batch of pictures taken over the same event may provide sufficient evidence for finding reliable correspondences.

Our second key assumption is that group-specific models ϕ^{g^*} give a highly accurate annotation: for example, the context-specific model ϕ^{rome} gives much more appropriate annotations for pictures from a visit to Rome than a general model ϕ trained on *all* Flickr pictures. This is due to two reasons: first, the *frequency* with which a tag appears varies between contexts (e.g., the term "skyscraper" should be expected more frequently in a "New York" context than in a "Rome" one). Second, the *appearance* of a tag varies (e.g., "buildings" in Rome look different to the ones in New York). Our approach will capture both kinds of information in group-specific distributions $P(t|g)$ (for tag frequency) and $P(v|t, g)$ (for appearance).

Our approach—as outlined so far—is independent of the kind of annotation model ϕ. In general, the proposed learning from Flickr groups could be integrated with a variety of approaches, simply by replacing a global training on all images with many group-specific ones. To demonstrate this, we will present two context-based extensions of different base models in the following.

C. Texton Model

We first introduce a context-based extension of a simple approach modeling class-conditional densities of patch occurrences in form of probability tables. The method resembles an early approach by Mori *et al.* [31] and has also been used as a baseline in image annotation before [11]. Similar to [11], we will refer to it as "texton model" in the following.

1) Texton Without Context: Let us first introduce the texton model in its standard form, i.e., when *not* using context and annotating each image independently. We infer the tag posterior for an image I by applying Bayes' rule:

$$\begin{aligned}\phi(\mathbf{x}', t) &\approx P(t|\mathbf{x}') \propto P(t) \cdot P(\mathbf{x}'|t)\\ &= P(t) \cdot \prod_{v=1}^{V} P(v|t)^{\mathbf{x}'_v}.\end{aligned} \tag{2}$$

The tag prior $P(t)$ and the visual word distribution $P(v|t)$ are both discrete probability tables learned from the annotations of the training set $\mathcal{X}$.

2) Texton With Context: To extend the texton model such that context is used, we replace the distributions $P(t)$ and $P(v|t)$ with group-specific equivalents $P(t|g)$ and $P(v|t, g)$.

Flickr Group Training: These group-specific distributions are trained by replacing the global training set $\mathcal{X}$ with a group-specific one $\mathcal{X}^g$. Let $n(\mathbf{x}) := \sum_{v=1}^{V} \mathbf{x}_v$ denote the number of patches in an image and $\mathbf{1}_c$ an indicator function evaluating to 1 iff. condition c is true (otherwise 0). Then we model

$$\begin{aligned}P(t|g) &= \frac{\sum\limits_{(\mathbf{x},\mathbf{t}) \in \mathcal{X}^g} \mathbf{1}_{t \in \mathbf{t}} \cdot n(\mathbf{x})}{\sum\limits_{(\mathbf{x},\mathbf{t}) \in \mathcal{X}^g} n(\mathbf{x})}\\ P(v|t, g) &= \frac{\sum\limits_{(\mathbf{x},\mathbf{t}) \in \mathcal{X}^g} \mathbf{1}_{t \in \mathbf{t}} \cdot \mathbf{x}_v}{\sum\limits_{(\mathbf{x},\mathbf{t}) \in \mathcal{X}^g} \mathbf{1}_{t \in \mathbf{t}} \cdot n(\mathbf{x})}.\end{aligned} \tag{3}$$

For practical reasons, we replace zero entries in these tables with small nonzero values.

Group Matching: A test batch $\mathcal{X}'$ is matched to a learned group using the maximum-likelihood approach described in (1):

$$\begin{aligned}g^* &= \arg\max_{g \in G} \; p(\mathcal{X}'|\phi^g)\\ &= \arg\max_{g \in G} \prod_{\mathbf{x} \in \mathcal{X}'} P(\mathbf{x}|g)\\ &= \arg\max_{g \in G} \prod_{\mathbf{x} \in \mathcal{X}'} \prod_{v=1}^{V} P(v|g)^{\mathbf{x}_v}\end{aligned} \tag{4}$$

whereas the group-specific distribution of visual words, $P(v|g)$, is estimated as

$$P(v|g) = \frac{\sum_{(\mathbf{x},\mathbf{t})\in\mathcal{X}^g} \mathbf{x}_v}{\sum_{(\mathbf{x},\mathbf{t})\in\mathcal{X}^g} n(\mathbf{x})}. \tag{5}$$

Group-Specific Annotation: Finally, each image $\mathbf{x}'$ is annotated according to (2); only $P(t)$ and $P(v|t)$ are replaced with $P(t|g^*)$ and $P(v|t,g^*)$.

D. PLSA Model

We also present a context-based extension of another model based on probabilistic latent semantic analysis (PLSA) [16]. PLSA belongs to the family of *topic models*, which—though originally introduced in the text domain [3], [16]—have frequently been used for image annotation [1], [11], [30]. For our context-based extension, we use a variant called "PLSA-Words" that has previously been presented by Monay and Gatica-Perez [30].

Like in Section III-C, each image I is represented by a bag-of-visual-words histogram $\mathbf{x}$ and a set of tags $\mathbf{t} \subset T$. Both tags and visual words are assumed to be generated by an image-specific mixture of a few latent aspects (or *topics*), which are denoted with $z \in \{1,\ldots,Z\}$ (the number of topics Z is assumed to be known and fixed). Visual words v and tags t are sampled from the following distributions:

$$\begin{aligned} P(t|I) &= \sum_{z=1}^{Z} P(t|z)\cdot P(z|I) \\ P(v|I) &= \sum_{z=1}^{Z} P(v|z)\cdot P(z|I). \end{aligned} \tag{6}$$

$P(z|I)$—the "topic mixture"—describes an image as a weighted combination of topics.

1) PLSA Without Context: We first briefly outline the standard PLSA-Words model when not using context (for more information, please refer to [30]). Both training and annotation are based on a maximization of the data likelihood, whereas optimization is carried out using expectation maximization (EM) or variants [16].

Training: In training, a set of annotated images $\mathcal{X}$ is used to learn the topic mixtures $P(z|I)$ for all images $I\in\mathcal{X}$, as well as topic-specific distributions $P(v|z)$, $P(t|z)$. First, the distribution of visual words is neglected, and the topic distribution $P(z|I)$ is learned by an EM maximization of the likelihood of the training images' *tags* only. Second, PLSA is run over the *visual words* to compute $P(v|z)$. Again, EM is used for optimizing the likelihood, but the topic distributions $P(z|I)$ learned in the previous step are kept fixed (see [30] for details).

Annotation: Given a previously unseen batch $\mathcal{X}'$ of test images I' to be labeled, each image $I'\in\mathcal{X}'$ is annotated independently. I' is represented by the feature $P(v|I') := \mathbf{x}'_\mathbf{v}/n(\mathbf{x}')$, and the tag distribution $P(t|I')$ is inferred in two steps: first—given $P(v|I')$—$P(z|I')$ is computed using a so-called "fold-in heuristic" [16]: the likelihood of $\mathbf{x}'$ is maximized using EM, whereas $P(v|z)$ (which was previously learned in training) is kept fixed. Second, the distribution of tags is estimated as

$$\phi(\mathbf{x}',t) \approx P(t|I') \approx \sum_{z=1}^{Z} P(t|z)\cdot P(z|I') \tag{7}$$

where $P(t|z)$ was learned previously in training.

2) PLSA With Context: Like for the texton model, the PLSA extension replaces the tag and visual word distributions $P(v|z)$ and $P(t|z)$ with group-specific equivalents $P(v|z,g)$ and $P(t|z,g)$ (again, g denotes a Flickr group like "Pets" or "Wedding Pictures"):

$$\begin{aligned} P(t|I,g) &\approx \sum_{z\in Z} P(t|z,g)\cdot P(z|I) \\ P(v|I,g) &\approx \sum_{z\in Z} P(v|z,g)\cdot P(z|I). \end{aligned} \tag{8}$$

This way, a single global annotation model is effectively replaced with many group-specific ones.

Flickr Group Training: A separate PLSA-based annotation model is learned for each Flickr group $g\in G$ on the group-specific training set $\mathcal{X}^g$. Similarly to the PLSA baseline, the EM algorithm is used; only the distributions $P(v|I)$ and $P(t|I)$ are replaced with group-specific equivalents $P(v|I,g)$ and $P(t|I,g)$.

Group Matching: Like for the texton model, a batch of test images $\mathcal{X}'$ is matched with a Flickr group using a maximum-likelihood criterion:

$$g^* = \arg\max_{g\in G} \prod_{\mathbf{I}'\in\mathcal{X}'} \prod_{v=1}^{V} P(v|I',g)^{\mathbf{x}'_\mathbf{v}} \tag{9}$$

with $P(v|I',g)$ from (8). Like for the texton model, the group is picked that maximizes the likelihood of observing $\mathcal{X}'$.

Group-Specific Annotation: Finally, annotation is carried out using (7), only that the specific model of group g^* is used:

$$\phi^{g^*}(\mathbf{x}',t) \approx P(t|I',g^*) \approx \sum_{z=1}^{Z} P(t|z,g^*)\cdot P(z|I'). \tag{10}$$

IV. Experiments

The key hypothesis of this paper is that image annotation can be improved significantly by matching a pictures' context to an appropriate pre-learned Flickr group and then employing a *context-specific* annotation.

This setup raises several key questions: 1) Does a context-specific annotation give improvements over standard single-image annotation? 2) How many pictures are required in a context to achieve such improvements? 2) Why do Flickr groups provide a good basis for learning visual contexts? 4) What is it that causes the improvements by a context-specific annotation—is it context-specific tag frequency or appearance information? Finally, 5) How does our approach compare to the state of the art? In the following, we answer these questions in a variety of quantitative experiments.

TABLE I
STATISTICS OF THE DATASETS USED IN SECTION IV

	tags	groups	images	batch size
Corel-Small	644	13	1,300	20
Corel-Big	1257	45	4,500	20
Corel-5K	374	50	5,000	10
Flickr-Small	544	8	8,000	20
Flickr-Big	252	609	83,406	20

A. Setup

We evaluate the proposed approach on several datasets. To allow comparison with other researchers' results, a first category of test cases has been sampled from the well-known *Corel Dataset* [32], a collection of photo groups constituting a well-known test case in image annotation benchmarking. Beyond this, we also tested our method on real-world content from Flickr, where visual contexts are directly learned from Flickr groups. In all cases, we learn contexts corresponding to different travel destinations and events (like "New York City Trip", "African Safari", and "Wedding Pictures"), simulating the annotation of users' personal photo collections. For an overview of the datasets used, please refer to Table I.

Corel-Small: We learn 13 contexts from 13 Corel folders (1300 pictures) corresponding to countries, regions, and cities (for example, "Africa" and "Kyoto"). These folders are used to simulate Flickr groups: from each folder, we use 80 pictures to train a group-specific annotation model, and the remaining 20 pictures are grouped to a batch for our context-specific annotation. A vocabulary of 644 Corel tags is used.

Corel-Big: A similar but larger dataset is compiled of 45 folders (4500 pictures), with a vocabulary of 1257 tags. Again, 20 images per folder were grouped to 45 test batches.

Corel-5K: We also want to give an impression of how the proposed context-based extension of image annotation compares to the state of the art. Therefore, our approach was evaluated on the well-known Corel-5K benchmark, a frequently used test case for image annotation with reference results for multiple standard methods [5], [9], [13], [20], [40]. The dataset consists of 50 folders (5000 images), whereas—according to standard practice—90 pictures per folder are used for training and 10 for testing. The standard tag vocabulary of 374 terms was used.

Flickr-Small: We downloaded 8000 images from eight Flickr groups related to travel destinations like "Rome" and "Maldives". To obtain a vocabulary of well-suited tags, the most frequent annotations in the dataset were filtered in two steps. First, a simple automatic vocabulary cleaning was done (removing stop words and numbers). Second, to improve the quality of tags further, unsuitable terms with little visual evidence (like "Olympus" or "great") were filtered manually, obtaining a vocabulary of 544 terms.

Flickr-Big: We also used a large-scale dataset drawn from a dense subset of the Flickr group space. We found 609 Flickr groups using 38 textual queries to the Flickr API related to a variety of topics, like "party", "nature", or "high dynamic range". The dataset contains 83 406 images (duplicates occurring in multiple groups were discarded, as they would overly simplify group matching and lead to biased results). Like for the Flickr-small set, a vocabulary of 252 terms was created by a refinement of the most frequent tags.

In all experiments, images are represented by bag-of-visual-words features obtained from a dense sampling of patches and their description using SURF [2] (Corel-Small, Corel-Big, Flickr-Small) or DCT coefficients [41] (Corel-5K). For the Flickr-Big dataset, the SiftGPU library[1] was used for a fast extraction of SIFT features on graphics hardware. In all tests, patches were clustered to codebooks of 2000 visual words trained on each specific dataset with a fast K-Means implementation [10]. The average number of patches per image was about 5000 (for Corel-Small, Corel-Big, Flickr-Small, and Corel-5K), or 500 (for Flickr-Big).

For the PLSA model (Section III-D), we present results for $Z = 20$ topics, which was found to give a good balance between speed and accuracy. For the texton model, we performed an additional dimensionality reduction of the visual features using PLSA [16], replacing high-dimensional visual word vectors with 64-dimensional topic posteriors.

As a performance measure, we use mean average precision (MAP) in all experiments (except for Corel-5K, where we refer to the standard measures of the benchmark): for each tag, the recall-precision curve is computed, and the average precision (i.e., the area under the curve) is averaged over all tags. Tags that do not appear in the test set were left out.

Finally, all experiments were repeated ten times with randomly re-sampled training and test data and average results over these ten runs are reported (except for Corel-5K, where standard training and test sets were used).

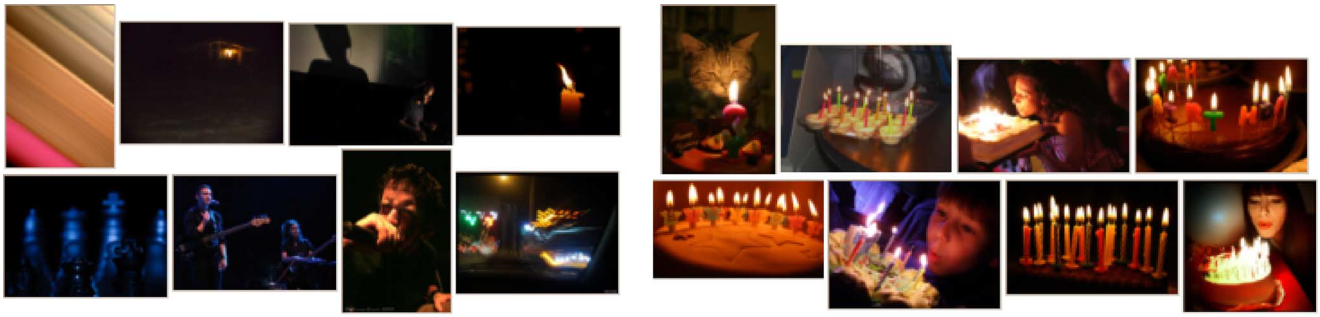
B. Results

1) Comparison With Single-Image Annotation: We first illustrate the effects of the proposed context modeling with a few sample results. Fig. 5 shows results for two test pictures from the Flickr-Small dataset. While the standard approach ("no context"), which annotates pictures individually with a global model, does not give satisfying results, the proposed approach ("with context") draws additional evidence from other pictures in the context—for example, the left picture is mapped to the "Paris" group, resulting in the correct tag "Eiffeltower".

Another result is illustrated in Fig. 4, where the eight pictures from the Flickr-Big test set with top scores for the tag "candle" are illustrated. Again, we see that results are strongly improved by modeling context: while the baseline system gives only a single correct match, the context-based approach yields eight hits (most of them were matched to a "birthday" Flickr group). These results indicate that we can improve image annotation significantly by modeling context from web categories.

Let us now quantitatively compare the proposed context-based approach with a separate annotation of individual images as that constitutes the state of the art. Results of this experiment are given in Fig. 6. The first—and most important—observation is that the proposed use of context ("with context", red) gives strong improvements over the standard annotation of individual images ("no context", yellow). These

[1] http://cs.unc.edu/~ccwu



(a) (b)

Fig. 4. Sample retrieval result: the top-scored images for the tag “candle” when using the texton model with (a) a plain annotation of individual images and (b) the proposed context-based model. While the base model gives only a single hit, the extension with context results in eight correct matches, all from “birthday” contexts.

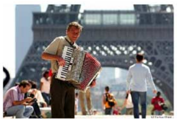
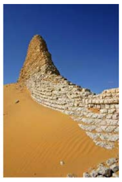
		
category	“Paris”	“Africa”
no context	**manhattan** **newyork**	**sea sunset**
with context	**eiffeltower**	**desert sand**

Fig. 5. Sample annotations using the PLSA model when performing a standard annotation with a single global model (top) and with the proposed context-based extension (bottom). While the standard approach gives incorrect annotations, our approach uses the *context* of pictures to map them to the right Flickr groups “Paris” (left) and “Africa” (right), and context-specific models achieve accurate annotations.

improvements are significant for both base models and over all datasets (paired t-test, level 99%), ranging from 109% (Corel-Small, texton model) to 497% (Flickr-Big, texton model). This result confirms our previous observations made in Figs. 4 and 5, and demonstrates that indeed a context-specific annotation provides more accurate results than labeling individual images.

Fig. 6 also shows a control run using a perfect group matching, i.e., a test batch is always matched to the correct context category (“context assigned”). We see that—compared to this control run—some performance loss occurs, which can be attributed to the unreliability of group matching based on visual features only.

2) Influence of Batch Size: Intuitively, we would expect that the more pictures we have in a context, the more reliably we could match them to a Flick group, and the better we would expect annotation to be. We confirm this hypothesis in another experiment, in which we vary the number of pictures in a test batch. Annotation models were trained on all datasets, leaving out 32 pictures per group for testing. These test pictures are aggregated to batches of size varied between 1, 2, 4, 8, 16, and 32. On each batch, context-based annotation was applied (we use both the texton and the PLSA version on different datasets). Results are averaged over 5–10 runs of random re-sampling, and plotted against the test batch size in Fig. 7. Fig. 7(a) shows the accuracy with which we correctly assign test batches to the

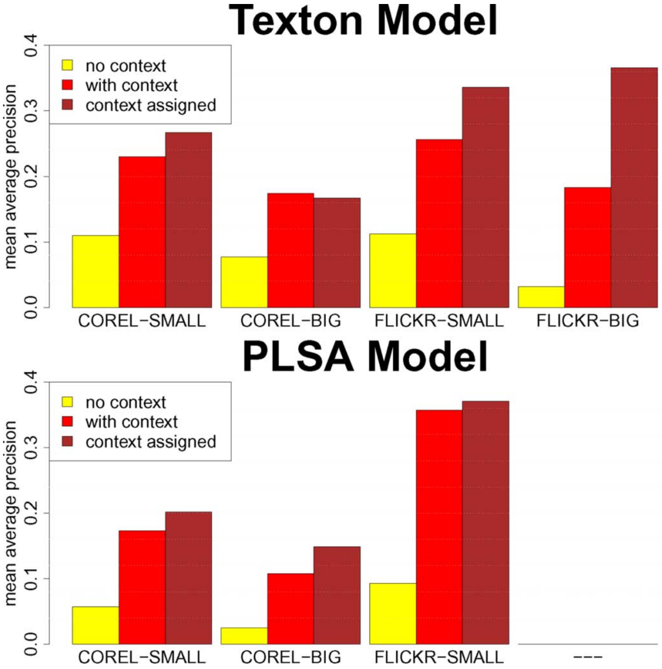
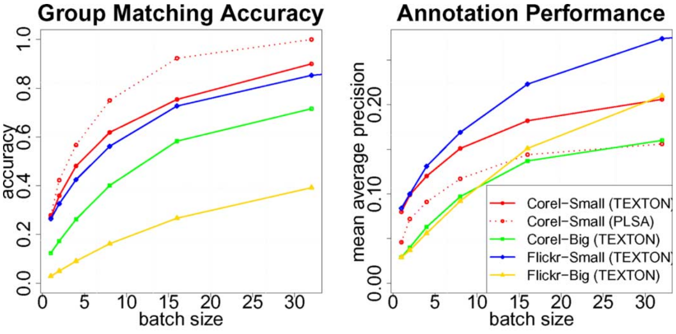


Fig. 6. Comparing the proposed context-based annotation (red) with the standard approach using a global model (yellow) and with a control run always using the correct group (brown). It can be seen that context modeling gives significant improvements.



(a) (b)

Fig. 7. Plotting (a) the accuracy of group matching and (b) annotation performance against the number of pictures in a batch. For a batch size of 1, our approach performs comparable to standard base models not using context. The more pictures we have in the context, the more reliable both group matching and overall annotation become.

TABLE II
COMPARING USER-DRIVEN CONTEXT CATEGORIES WITH GROUPS LEARNED BY AN UNSUPERVISED CLUSTERING

mean avg. prec.	visual clustering		tag-based clustering		**user-defined**
	K-Means	PLSA	K-Means	PLSA	
Corel-Small	0.081	0.081	0.067	0.072	**0.230**
Corel-Big	0.034	0.034	0.024	0.027	**0.167**
Flickr-Small	0.091	0.093	0.081	0.069	**0.256**

category they have been sampled from. Fig. 7(b) illustrates the resulting annotation performance.

We see that for batchsize 1 (i.e., each image is mapped to a context category individually), the accuracy of group matching is low, and correspondingly image annotation is inaccurate (results are comparable to when using the "no context" base models in Fig. 6). However, when increasing the number of pictures in a context, a more reliable group matching becomes possible, and correspondingly tagging results improve distinctly. A paired t-test (level 99%) indicates that these improvements are statistically significant for a batch size of at least 2 (Corel-Small, Flickr-Small, Flickr-Big) or eight images (Corel-Big), respectively. This indicates that in general, as many pictures as possible in a context are desirable, but that improvements can already be achieved for small groups.

3) Why Flickr Groups?: We have argued in Section III-A that the user-driven categorization and richness of Flickr groups are essential for a successful learning of context categories. In the following, we present experiments validating these hypotheses.

User-Driven Categorization: By using Flickr group information, our approach implicitly employs the fact that users on the web manually sort their pictures into semantic categories. This manual effort made by a large web-based community leads to a meaningful structure where pictures in the same group are usually semantically related. However, one might argue that supervised learning from this structure could be substituted with an unsupervised clustering of training pictures. We compare such an automatic grouping with a direct learning from the Flickr group structure: training sets are clustered into a number of partitions equal to the number of user-generated categories. Then, the proposed context-based approach is followed, where style-coherent batches of test images are matched with these clusters instead of Flickr groups.

Results of this experiment are displayed in Table II. We tested K-Means and PLSA clusterings based on the visual representation of training images as well as on tag information (using a bag-of-words representation). It can be seen that such an unsupervised learning of context categories comes with a performance loss of 65%–85% compared to learning from the Flickr group structure. This shows that an unsupervised grouping poses a difficult challenge, such that context learning requires human users as a corrective for semantic categorization.

Richness of Groups: A second key benefit is that Flickr groups provide an extremely rich collection of over 200 000 very specific categories. To demonstrate that this richness is vital for generalizing to a specific user's content, we apply our context-based approach to images *not* drawn from any learned category. This is done by using the same test batches as before, but now explicitly prohibiting each test batch to match the category it has been drawn from. For example, we try to explain pictures from a "Rome" category by models for "New York" pictures, "Safari" pictures, "Greece" pictures, etc. This setup (called "cross-group") was applied to all our test datasets using the texton model. We compare it with a standard baseline annotating images individually ("no context").

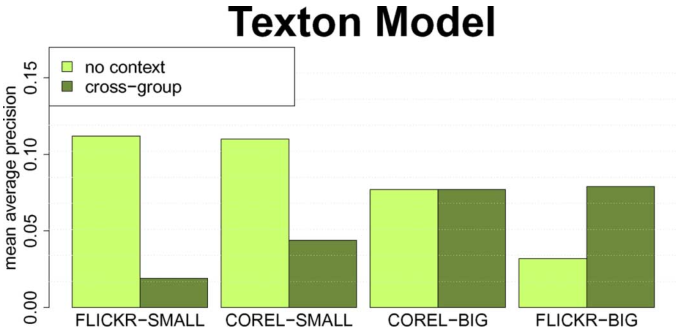


Fig. 8. Cross-group generalization experiment: datasets are ordered with increasing number of groups from left to right. The more groups we use for learning, the better we can generalize to specific content unseen in training, and the better the context-based approach performs relative to the "no-context" baseline.

Results are illustrated in Fig. 8. The datasets are ordered by the number of categories, from Flickr-Small (8) over Corel-Small (13) and Corel-Big (45) to Flickr-Big (608). We observe a clear trend with respect to the number of categories: for the datasets with few groups, generalizing to new content is difficult, and the context-based approach performs significantly worse than the standard baseline. However, the more categories we employ, the better we can explain batches from groups unseen in training. Correspondingly, annotation performance improves: at 45 categories, the context-based approach is on par with the baseline, and when learning from the full richness of Flickr groups, we observe that context gives significant improvements (from 3.2% to 7.9%) even without allowing to map test pictures to the Flickr group to which they belong. The latter would correspond to a MAP of 18.3%. Obviously, the richer the category space we employ for visual learning, the better we can explain batches of content that do *not* exactly match the categories learned.

We also examined the cross-group runs on the Flickr-Big dataset in depth by manually categorizing 1000 group matches into four categories (the decision was based on the group name).

1) **Exact**: the test content is matched exactly to the Flickr group it comes from.
2) **Appropriate**: the test content is matched to a group that can be expected to give an appropriate annotation model. Examples are "Best of Cats" → "Cat Lovers" and "Wedding Planners" → "Inspirational Wedding Photography".
3) **Related**: the test content and the matched group are widely related. Examples are "Hiking in Canada" → "Sights of Nature" or "Beijing 2008 Olympics" → "2006 FIFA World Cup".
4) **No Match**: the matched group cannot be expected to explain the test content well. Examples are "War Photo

TABLE III

ACCURACY AND SPEED OF GROUP MATCHING ON THE FLICKR-BIG DATASET. **LEFT**: IN MANY CASES, AN APPROPRIATE OR RELATED FLICKR GROUP IS PICKED FROM THE 609 ONES LEARNED, EVEN WHEN ENFORCING GENERALIZATION TO NEW GROUPS ("CROSS-GROUP"). **RIGHT**: THE AVERAGE COST OF GROUP MATCHING FOR BATCHES OF 20 IMAGES FOR THE TEXTON MODEL AND THE PLSA MODEL (*ESTIMATED)

	Accuracy (%)			
category	1 (exact)	2 (appr.)	3 (rel.)	4 (none)
with-group	32.9	24.2	13.6	29.3
cross-group	0.0	22.2	30.2	47.4

Speed (sec.)	
model	time
texton	0.05
PLSA	2476.60*

Journalism" → "Alaska Birds" or "Wedding Planners" → "Board Games".

Results are illustrated in Table III. We see that—if allowing test pictures to match the Flickr group they are drawn from ("with-group")—in more than two thirds of the cases, the system picks at least a widely related match out of the 608 training groups. For the "cross-group" systems, matching becomes more difficult as we allow no more exact hits. However, our approach still finds a related group in more than 50% of cases. Overall, these results indicate that Flickr groups—with their user-driven categorization and richness of group space—form an excellent basis for a visual learning of contexts.

4) Influence of Tag Information and Appearance: As indicated previously, one would expect that improvements by a context-specific annotation are based on both *frequency* and *appearance*. Our approach employs both these kinds of information by modeling group-specific distributions $P(t|g)$ and $P(v|t,g)$, which give more accurate models for a tags' frequency and appearance than $P(t)$ and $P(v|t)$.

In the following, we evaluate either of these information sources separately by comparing the proposed approach ("full model") with two baselines.

- **tags-only**: The system uses visual information to map a batch of pictures to a learned group, and then only the group-specific tag prior $P(t|g^*)$ is employed as a score (i.e., visual content is only used for group matching but not for annotation).
- **appearance-only**: After mapping a batch of pictures to a learned Flickr group, the system uses a group-specific appearance model $P(v|t,g)$ but a global tag prior $P(t)$. For details on how this is realized for the PLSA model, please refer to our previous publication [8].

Our results in Fig. 9 compare the full approach with these two strategies making limited use of context. We see that combining a context-specific appearance and tag occurrence model outperforms both baselines, which indicates that—for a context-based approach to work—it is beneficial to make use of both context-dependent appearance and tag information.

C. Corel-5K Benchmark

To get an impression of how the proposed context learning compares to the state of the art in image annotation, we perform an evaluation on the well-known Corel-5K benchmark. We report the same performance measures as in the literature: for each image, the top five annotations are returned, and for each tag the per-word precision and per-word recall are measured over all test images. These values were averaged over all 251 tags occurring in the test set, and also combined to an F-measure. Finally, also the number of tags t with a recall greater than zero was recorded.

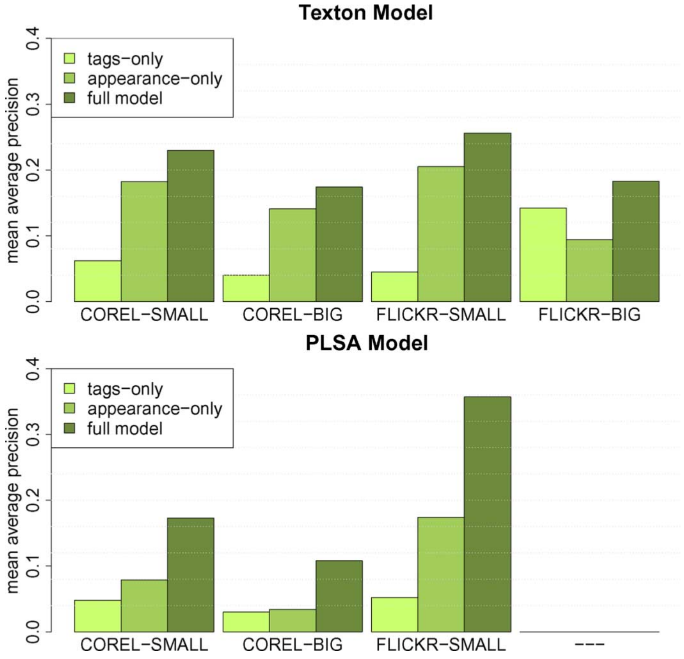


Fig. 9. Comparing our model (which uses group-specific tag occurrence *and* appearance statistics) with two baselines exploiting only one of both information sources. It can be seen that combining tag and appearance information is vital for a successful context-specific annotation (the PLSA model was not applied to the Flickr-Big set due to high computational cost, compare Table III).

TABLE IV

RESULTS ON THE COREL-5K BENCHMARK. BY MAKING USE OF CONTEXT INFORMATION, A COMPETITIVE PERFORMANCE COMPARED TO THE STATE OF THE ART IS ACHIEVED

Approach	words rec.>0	avg. prec.	avg. rec.	F-measure
co-occurrence [31]	19	0.02	0.03	0.02
Translation [9]	49	0.04	0.06	0.05
kernel densities (co-occ.) [6]	91	0.11	0.13	0.12
SVDCos [40]	102	0.15	0.15	0.15
CRM [20]	107	0.16	0.19	0.17
kernel densities (CT-3x3) [48]	114	0.18	0.21	0.19
InfNet [29]	112	0.17	0.24	0.20
CSD-Prop [40]	130	0.20	0.27	0.23
MBRM [13]	122	0.24	0.25	0.24
SML [5]	137	0.23	0.29	0.26
CSD-SVM [40]	127	0.25	0.28	0.26
JEC [27]	139	0.27	0.32	0.29
PicSOM [43]	**—**	**0.35**	**0.35**	**0.35**
PLSA (no context) [30]	57	0.04	0.09	0.05
PLSA (with context)	**141**	**0.25**	**0.39**	**0.31**
texton (no context)	26	0.04	0.05	0.04
texton (with context)	**119**	**0.21**	**0.35**	**0.26**

Quantitative results are illustrated in Table IV, including figures reported by other researchers: the co-occurrence model by Mori *et al.* [31], the machine translation model by Duygulu *et al.* [9], two relevance models by Manmatha and co-workers [13], [20], supervised multi-class labeling by Carneiro *et al.* [5], kernel densities [6], [48], feature-centric approaches [27], [43], and several other annotation models [29], [40]. We also included the texton and PLSA models when applied to individual images ("no context").

Our results show that—while both models perform poorly when annotating images individually—both context-based extensions lead to significant improvements and achieve a competitive performance. For example, the context-based PLSA model gives a recall of 39%, precision of 25%, and F-measure of 31%, which is comparable to the best systems reported to date [43].

It is important to keep in mind that—compared to the other methods in Table IV—our approach requires some extra information: essentially, we assume that a user provides not only pictures to be annotated but also tells us which of them belong to a coherent batch. Correspondingly, we do not claim a better annotation model compared to others. Instead, our intention is to demonstrate that context—*if* it is available as an additional information source—has the potential to drive image annotation to a new quality level. Table IV demonstrates this for our two test models, and similar improvements might be possible for others. For example, one candidate is the PicSOM model [43], which gives the best results reported to date on Corel-5K. While we have employed visual words as a standard approach in this paper, PicSOM draws its strength from a combination of ten diverse types of features. An extension of our approach with more features only requires to adapt our strategy for matching batches of pictures for this specific feature set.

V. Discussion—Application to Web-Scale Group Spaces

While our evaluation in Section IV has included datasets of up to 80 000 images and over 600 Flickr groups, web-based portals offer much larger datasets (in Flickr, there are more than 200 000 groups). What is to be expected when applying the proposed visual learning at such scale?

On the one hand, we have provided experimental evidence that rich web-scale group space might significantly improve the generalization to a particular users' photos (see Fig. 8). On the other hand, scalability issues arise, as the time effort of group matching and storage requirements grow linearly with the number of groups. Still, for simple annotation models, our approach remains applicable at large scale: for example, group matching in the texton model merely demands one scalar product computation per group (4), which takes only 0.05 s on the 609 groups of the Flickr-big set (see Table III). Taking into account that this is required only once per batch (not per image) and that our approach is easily parallelizable, it can be concluded that for simple annotation models, large-scale group-specific annotation is possible.

The situation is different when it comes to more complex annotation approaches like the PLSA model. Here, the more costly group matching strategy requires a complete EM optimization per group, which is too time consuming for very large group spaces (see Table III). On the other hand, more complex annotation models may offer benefits in terms of accuracy: for example, see Fig. 6, where PLSA accuracy suffers less from automatic group matching on the Flickr-small set, or Table IV, where PLSA outperforms the texton model on the Corel-5K benchmark. To benefit from these strengths and apply visual learning from Flickr groups with more complex base models, an interesting alternative might be to aggregate Flickr groups to higher-level categories (see the promising work on "Flickr hypergroups" [34] by Negoescu *et al.*). With these approaches, the diversity of the Flickr group space may be preserved while keeping the number of categories at a manageable size.

Finally, another issue in practice is correlations between Flickr groups, which may address similar topics or even share the same images. We removed duplicate pictures in our experiments to avoid biased results. In a practical large-scale setting, however, group correlation might provide valuable extra information: shared images may provide strong hints for relations between categories, which may help to form higher-level "hypergroups". One can also envision an approach that maps image batches to *sets* of related groups and combines their group-specific annotation results.

VI. Conclusion

We have suggested a novel extension to image annotation that employs web-based user-driven category information like Flickr groups as an additional information source. Our approach assumes images to come with a *context* of related pictures (e.g., taken over the same event). This context is matched with Flickr groups, and then a group-specific annotation is applied. Significant improvements of up to 100% and more have been validated on samples from the Corel dataset as well as real-world Flickr data. We have also analyzed the validity of Flickr groups as a basis for our approach, and have shown two key characteristics they offer for learning visual contexts, namely a user-driven categorization and a rich group space, which aids in generalizing to novel categories.

As the proposed approach can be applied as a wrapper around a variety of base models, one promising direction for future work will be the integration with other annotation approaches. In this paper, extensions for two generative models have already been presented, namely PLSA and a texton model. Whereas the latter comes with a highly efficient group matching applicable at large scale, the former offers better accuracy for small-to-medium size datasets. As these models are quite different, our results indicate that other annotation approaches like nearest neighbor techniques or discriminative classifiers will likely benefit as well.

Another interesting issue is the exploration of higher-level context categories, which may improve both the scalability and accuracy of the proposed approach further: while Flickr groups offer very specific models, the number of training samples per group is rather limited (usually a few hundred pictures). It remains to be investigated whether coarser categories (e.g., higher-level "Flickr hypergroups" [34]) allow us to learn models that are less specific but founded on a more solid basis of training samples, and thus even more successful.

Acknowledgment

The authors would like to thank R. Manmatha and S. Feng for kindly providing information on the Corel-5K dataset.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[2] H. Bay, T. Tuytelaars, and L. van Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[3] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[4] L. Cao, J. Luo, H. Kautz, and T. Huang, "Annotating collections of photos using hierarchical event and scene models," in *Proc. CVPR*, 2008, pp. 1–8.

[5] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[6] A. Llorente Coto and S. Rüger, "Can a probabilistic image annotation system be improved using a co-occurrence approach?," in *Proc. SAMT Workshop Cross-Media Information Analysis and Retrieval*, 2008.

[7] M. Cristani, A. Perina, U. Castellani, and V. Murino, "Geo-located image analysis using latent representations," in *Proc. CVPR*, 2008.

[8] M. Duan, A. Ulges, T. Breuel, and X.-Q. Wu, "Style modeling for tagging personal photo collections," in *Proc. CIVR*, 2009.

[9] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. ECCV*, 2002, pp. 97–112.

[10] C. Elkan, "Using the triangle inequality to accelerate KMeans," in *Proc. ICML*, 2003, pp. 147–153.

[11] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. CVPR*, 2005, pp. 524–531.

[12] S. Feng and R. Manmatha, "A discrete direct retrieval model for image and video retrieval," in *Proc. CIVR*, 2008, pp. 427–436.

[13] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. CVPR*, 2004.

[14] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," *Comput. Vis.*, vol. 2, pp. 1816–1823, 2005.

[15] A. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen, "Image annotation using personal calendars as context," in *Proc. ACM MM*, 2008, pp. 681–684.

[16] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, pp. 177–196, 2001.

[17] A. Hrishikesh, G. Toderici, and J. Yagnik, "Video2Text: Learning to annotate video content," in *Proc. Workshop Int. Multimedia Mining*, 2009.

[18] J. Jeon and R. Manmatha, "Using maximum entropy for automatic image tagging," in *Proc. CIVR*, Jul. 2004, pp. 24–32.

[19] H. Kück, P. Carbonetto, and N. de Freitas, "A constrained semi-supervised learning approach to data association," in *Proc. ECCV*, 2004, pp. 1–12.

[20] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press, 2004.

[21] J. Leon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. Int. Conf. Research and Development in Information Retrieval*, 2003, pp. 119–126.

[22] J. Li and J. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.

[23] L.-J. Li, G. Wang, and L. Fei-Fei, "OPTIMOL: Automatic object picture collection via incremental model learning," in *Proc. CVPR*, 2007.

[24] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma, "Image annotation by large-scale content-based image retrieval," in *Proc. ACM MM*, 2006.

[25] X. Li, C. Snoek, and M. Worring, "Annotating images by harnessing worldwide user-tagged photos," in *Proc. ICASSP*, 2009.

[26] X. Li, C. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[27] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. ECCV*, 2008, pp. 316–329.

[28] H. M. Castro, E. Sucar, and E. Morales, "Automatic image annotation using a semi-supervised ensemble of classifiers," in *Proc. Iberoamerican Congr. Pattern Recognition*, 2007, pp. 487–495.

[29] D. Metzler and R. Manmatha, "An inference network approach to image retrieval," in *Proc. CIVR*, 2004, pp. 42–50.

[30] F. Monay and D. Gatica-Perez, "PLSA-based image annotation: Constraining the latent space," in *Proc. ACM MM*, 2004, pp. 348–351.

[31] Y. Mori, T. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. Workshop Multimedia Intelligence Storage and Retrieval Management*, 1999.

[32] H. Müller, S. Marchand-Maillet, and T. Pun, "The truth about Corel – evaluation in image retrieval," in *Proc. CIVR*, 2002, pp. 38–49.

[33] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. Joint Conf. Digital Libraries*, 2005, pp. 178–187.

[34] R. Negoescu, B. Adams, D. Phung, S. Venkatesh, and D. Gatica-Perez, "Flickr hypergroups," in *Proc. ACM MM*, 2009.

[35] R. Negoescu and D. Gatica-Perez, "Analyzing Flickr groups," in *Proc. CIVR*, Jul. 2008, pp. 417–426.

[36] M. Renn, J. van Beusekom, D. Keysers, and T. Breuel, "Automatic image tagging using community-driven online image databases," in *Proc. Int. Workshop Adaptive Multimedia Retrieval*, 2008.

[37] P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 88–98, Jan. 2005.

[38] A. Setz and C. Snoek, "Can social tagged images aid concept-based video search?," in *Proc. ICME*, 2009, pp. 1460–1463.

[39] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*. New York: Springer-Verlag, 2006, pp. 127–144.

[40] J. Tang and P. Lewis, "A study of quality issues for image auto-annotation with the Corel dataset," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 384–389, Mar. 2007.

[41] A. Ulges, C. Schulze, D. Keysers, and T. Breuel, "A system that learns to tag videos by watching Youtube," in *Proc. ICVS*, 2008.

[42] A. Ulges, C. Schulze, M. Koch, and T. Breuel, "Learning automatic concept detectors from online video," *Comp. Vis. Image Understand.*, vol. 114, pp. 429–438, 2010.

[43] V. Viitaniemi and J. Laaksonen, "Empirical investigations on benchmark tasks for automatic image annotation," in *Proc. VISUAL*, 2007.

[44] C. Wang, F. Jing, L. Zhang, and H. Zhang, "Image annotation refinement using random walk with restarts," in *Proc. ACM MM*, 2006.

[45] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[46] C. Yang, M. Dong, and F. Fotouhi, "Region based image annotation through multiple-instance learning," in *Proc. ACM MM*, 2005.

[47] Q. Yang, X. Chen, and G. Wang, "Web 2.0 dictionary," in *Proc. Int. Conf. Image and Video Retrieval*, 2008, pp. 591–600.

[48] A. Yavlinsky, E. Schofield, and S. Rüger, "Automated image annotation using global features and robust nonparametric density estimation," in *Proc. CIVR*, Jul. 2005, pp. 507–517.

Adrian Ulges received the diploma degree in computer science (with honors) and the Ph.D. degree in computer science from the University of Kaiserslautern, Kaiserslautern, Germany, in 2005 and 2009, respectively.

He is currently a Senior Researcher with the Multimedia Analysis and Data Mining Group at the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern. His research interests are in pattern recognition and multimedia analysis, with a particular focus on visual learning from web information. He has published over 25 scientific papers in the areas of content-based image and video retrieval, multimedia forensics, and document image analysis.

Dr. Ulges has been active as a reviewer and program committee member, and has been co-chair of the International Workshop on Multimedia in Forensics, Security and Intelligence (MiFor) 2010.

Marcel Worring (M'03) received the M.Sc. degree (honors) in computer science from the VU Amsterdam, Amsterdam, The Netherlands, in 1988 and the Ph.D. degree in computer science from the University of Amsterdam in 1993.

He is currently an Associate Professor in the Intelligent Systems Lab Amsterdam of the University of Amsterdam. His research focus is multimedia analytics, the integration of multimedia analysis, multimedia mining, information visualization, and multimedia interaction into a coherent framework yielding more than its constituent components. With the MediaMill team, he develops techniques for semantic video indexing and interactive exploration of large video archives which have been successful over the last years in the TRECVID benchmark, the de-facto standard on the topic. He has published over 100 scientific papers covering a broad range of topics from low-level image and video analysis up to multimedia analytics.

Dr. Worring was co-chair of the 2007 ACM International Conference on Image and Video Retrieval in Amsterdam, and co-initiator and organizer of the VideOlympics. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and of the *Pattern Analysis and Applications* journal.

Thomas Breuel (M'08) received degrees from the Massachusetts Institute of Technology and Harvard University, both in Cambridge, MA.

He is a Professor of computer science and Head of the Image Understanding and Pattern Recognition (IUPR) research group at the University of Kaiserslautern, Kaiserslautern, Germany and a consultant in Palo Alto, CA. His research group works in the areas of image understanding, document imaging, computer vision, and pattern recognition. Previously, he was a researcher at Xerox PARC, the IBM Almaden Research Center, IDIAP, Switzerland, as well as a consultant to the U.S. Bureau of the Census.