Style Modeling for Tagging Personal Photo Collections

Manni Duan
Dept. of EEIS, USTC
Hefei, China
mnduan@mail.ustc.edu.cn

Adrian Ulges
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany
ulges@iupr.dfki.de

Thomas M. Breuel
DFKI and TU Kaiserslautern
Kaiserslautern, Germany
tmb@iupr.dfki.de

Xiu-qing Wu
Dept. of EEIS, USTC
Hefei, China
wuxq@mail.ustc.edu.cn

ABSTRACT

While current image annotation methods treat each input image individually, users in practice tend to take multiple pictures at the same location, with the same setup, or over the same trip, such that the images to be labeled come in groups sharing a coherent "style".

We present an approach for annotating such style-consistent batches of pictures. The method is inspired by previous work in handwriting recognition and models style as a latent random variable. For each style, a separate image annotation model is learned. When annotating a batch of images, style is inferred using maximum likelihood over the whole batch, and the style-specific model is used for an accurate tagging.

In quantitative experiments on the COREL dataset and real-world photo stock downloaded from Flickr, we demonstrate that – by making use of the additional information that images come in style-consistent groups – our approach outperforms several baselines that tag images individually. Relative performance improvements of up to 80% are achieved, and on the COREL-5K benchmark the proposed method gives a mean recall/precision of 39%/25%, which is the best result reported to date.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Indexing

General Terms

Algorithms, Measurement, Experimentation

Keywords

Image Annotation, Style Modeling

1. INTRODUCTION

Image annotation is targeted at automatically labeling pictures with semantic "tags", which are associated with objects in the image, locations, scene types, or activities. The task poses a difficult challenge due to enormous intra-class variation and large vocabularies of potential tags, and current systems do not give a performance sufficient for a fully automatic labeling. Yet, image annotation is useful in applications like semi-automatic tagging or search.

Figure 1: An illustration of style consistency for image annotation: when viewed individually, the bottom right image is difficult to annotate automatically (potential tags might be "forest" or "park"). By taking into account the fact that other pictures in the same batch show mostly urban scenes, this conflict can be disambiguated and the correct tag is identified to be "park".

Typically, tagging systems learn statistical models of concept appearance and use them to label previously unseen pictures, dealing with each image individually. In practice, however, the pictures we take tend to come in groups – for example, imagine categories at social photo sharing websites in which pictures are organized [21], or consider a user coming back from a holiday trip and annotating a batch of pictures he took there. In both cases, the images in a group are not independent (as is assumed by most annotation models) but are correlated, sharing similar locations and capture conditions.

We focus on this situation where the input images to be annotated come in correlated groups. Two basic assumptions are made:

1. A grouping of pictures is assumed to be given, i.e. it is known which images belong to a batch. Users may provide this information explicitly (by grouping pictures previous to annotation) or implicitly (by placing them in the same folder or

CIVR'09, July 8–10, 2009, Santorini, GR.

uploading them to the same Flickr group). Also, grouping information may be inferred automatically from meta-data like capture time and location.

2. The images in a batch are assumed to share a certain coherent appearance. The reasons for such *style consistency* may be manifold and subtle: for example, the pictures in Figure 1 could be considered to form a style-consistent batch as they have been taken on the same trip to Rome and thus show similar buildings, weather conditions, and shot types.

In the following, we will assume that images in a batch are *isogenous* (i.e. sampled from the same source [23]), and that there is a finite number $s_1, .., s_m$ of such sources (or *styles*). Our use case will be the tagging of personal holiday snapshots, where different styles correspond to different kinds of holiday trips. For example, there might be a "sightseeing" style s_1 and a "nature trip" style s_2. Both are considered generative sources from which a variety of images is sampled with a certain distribution – for example, pictures from the "nature trip" style tend to show close-ups of animals as well as panoramic landscape views, while pictures from the "sightseeing" style show street scenes and snapshots of buildings.

We demonstrate in the following that – if the images to be annotated come in style-consistent groups – taking style information into account can give significant improvements over a plain annotation of individual images. This idea is illustrated in Figure 1: consider the image at the bottom right showing an outdoor scene with trees and greenery. Using evidence from this single image only, an automatic annotation system might easily confuse the tags "forest" and "park". However, if further taking into account that the image belongs to a batch showing mostly urban scenes, this ambiguity can be resolved, and the correct tag "park" can be inferred.

To achieve such disambiguation, we turn to statistical models from handwriting recognition. This domain bears a strong resemblance with the annotation of style-consistent image batches: like groups of pictures, handwritten documents can be viewed as collections of samples (namely the single characters) sharing a consistent style. In both cases, the concept of style is merely driven by sample appearance, and the reasons for different styles can be manifold and subtle. Yet, it has been demonstrated for handwriting – and will be in the following for image annotation – that style can be captured effectively. For this purpose, we combine image annotation with a style model from Sarkar and Nagy [23]. Style is modeled as a latent random variable, and for each style a specific annotation model is built. Generally, a variety of probabilistic image annotation methods could be used for this purpose (though for this paper we choose the approach by Monay and Gatica-Perez [18] based on probabilistic latent semantic analysis (PLSA) [12]). When tagging a batch of images, a reliable style decision is made by maximizing the likelihood of the whole batch, and tags for each image are inferred using the accurate style-specific model.

We test our approach in experiments on the COREL dataset and real-world photo stock downloaded from Flickr. As our focus is on personal holiday snapshots, we choose our test styles to be location- or event-based (though our model is not restricted to those cases in general). For the COREL data, a 1:1 correspondence between picture batches and COREL folders (like "England" and "Kyoto") is imposed. Similarly, in the Flickr case style-coherent batches correspond to Flickr groups (we distinguish travel scenarios like "African Safari" and "New York Sightseeing"). It is demonstrated that style-consistency helps image annotation to disambiguate and improves the overall tagging performance significantly compared to an image-wise annotation.

2. RELATED WORK

Since Mori et al's pioneering work on automatic image annotation [19] a variety of approaches have been suggested. Usually, an image $\mathcal{X}$ is viewed as a collections of local image regions $\mathcal{X} = \{x_1, .., x_n\}$, which can be obtained using a segmentation (like in [15]) or a sampling of local patches [10]. The goal is to map the image to tags t from a pre-defined vocabulary T. Since tags of interest can include all kinds of semantic concepts, image annotation unifies related tasks such as object category recognition [9] and scene recognition [22].

The most frequent approach is to infer the posterior $P(t|\mathcal{X})$ by modeling a joint distribution of local features and tags $P(x, t)$ (some exceptions based on global image similarity exist, like [16]). For this joint distribution, topic models have been suggested that mine collections of annotated training pictures for latent aspects [2, 18]. Other options are Gaussian mixtures [4] or relevance models [10, 15]. Approaches from a second category view image annotation as a weakly labeled learning problem: it is assumed that the presence of a tag is caused by a certain region in the image, and tagging involves the explicit identification of "relevant" regions. To infer this mapping between regions and tags, the EM algorithm has been suggested [7] as well as multiple instance learning techniques [26] or probabilistic models formulating constrained learning problems [13]. While all these methods differ in terms of features and underlying statistical models, the approach presented here is orthogonal to those distinctions. Instead, our approach is targeted at exploiting information that is present in style-consistent *batches* of pictures. The proposed method can generally be used as a wrapper around probabilistic image annotation models.

More recently, a number of image annotation and categorization methods have been proposed which employ structural information in image collections in a fashion similar to the proposed method. The most frequent approach is to group the pictures of personal photo collections to *events* that resemble the notion of different *styles* in our approach. Cao et al. [3] use the resulting grouping information in form of correlation terms in a conditional random field (CRF) model. Naaman et al. [20] propagate labels through the event structure to estimate the identity of persons in picture collections. Gallagher et al. [11] match images with events in a calendar, whereas a content-based categorization serves as a filter. Finally, Cristani et al. [6] present an extension of a topic model that integrates the capture location of pictures as a latent aspect. All these approaches demonstrate impressive results by employing the structure of image collections beyond an annotation of individual images. In this paper, we follow the same target, but approach the problem from a slightly different perspective: the approaches mentioned above focus strongly on the challenge of how to infer a grouping of images using meta-data such as capture times [11] or capture locations [6]. While this is clearly an important problem, we focus on situations in which such a grouping is already given. For these situations, we present a well-justified model for image annotation and validate performance improvements over tagging individual images.

Finally, as our approach is adopted from handwriting recognition, related work in this domain on the classification of style-consistent batches should be mentioned. Baird and Nagy [1] present an approach for font adaptation based on the assumption of style consistency. Their approach adapts a given multi-font baseline classifier in a iterative self-training. Sarkar and Nagy [23] model style as a finite random variable that is inferred using a maximum-likelihood approach over the whole batch, an approach that we adopt here. Beyond this, hierarchical Bayesian methods have been proposed [17] that model style implicitly in form of

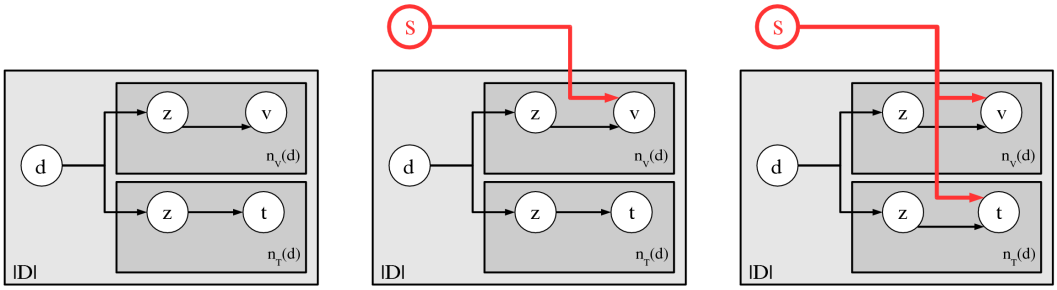


Figure 2: Graphical models depicting the sample generation process for image features v and tags t. (a) the baseline PLSA model. (b) the "appearance-only" style extension. (c) the "appearance-and-tags" style extension.

prior distributions on the underlying parameters of a sample generation process. Since these methods overcome the need for explicit discrete styles, they might be an interesting extension of the approach presented in this paper.

3. APPROACH

In this section, a framework is presented that integrates a PLSA-based image annotation model [18] with a style model adapted from handwriting recognition [23] to achieve an improved annotation of style-consistent image batches. We start with the plain annotation model, which will also serve as a baseline in later experiments (Section 3.1). After this, two style-based extensions of this model will be presented (Sections 3.2 and 3.3). A graphical illustration of all models is given in Figure 2.

3.1 Baseline: Coupled PLSA

This section briefly introduces the image annotation method by Monay and Gatica-Perez [18], which models two distributions coupled in one PLSA model (one for image tags and one for discretized image features called "visual words" [24]). A set of training images D is given, whereas each image d is represented by tags $t \in T$ and a set of visual words v from a vocabulary V. The PLSA model posits that tags and visual words are conditionally independent given latent *topics* $z \in Z$. Both are sampled from the following distributions by marginalizing over topics:

$$\begin{aligned} P(v|d) &= \sum_{z \in Z} P(v|z) \cdot P(z|d) \\ P(t|d) &= \sum_{z \in Z} P(t|z) \cdot P(z|d) \end{aligned} \tag{1}$$

The distribution $P(z|d)$ assigns topics to images, and the topic distributions $P(t|z)$ and $p(v|z)$ determine how tags and visual words are sampled from each topic. The number of topics $|Z|$ is assumed known and fixed. Both learning and inference are based on a maximization of the following likelihood:

$$\mathcal{L} = \prod_{d \in D} \left[P(d) \cdot \prod_{v} P(v|d)^{n(v,d)} \cdot \prod_{t} P(t|d)^{n(t,d)} \right], \tag{2}$$

where $n(., d)$ denotes the number of occurrences of tags and visual words in image d.

3.1.1 *Learning*

For PLSA models, learning (i.e., the estimation of topics and topic distributions) is based on expectation maximization (EM) or variants [12]. For the coupled PLSA model used here, we follow a similar asymmetric two-step procedure for which Monay and Gatica-Perez have reported improved results [18]:

1. The distribution of visual words is neglected, and the topic distribution $P(z|d)$ is learned by maximizing the likelihood of the textual image descriptions only:

$$\mathcal{L}_T = \prod_{d \in D} \left[P(d) \cdot \prod_{t} P(t|d)^{n(t,d)} \right]. \tag{3}$$

   For optimization, expectation maximization is used (for more information, please refer to [12]).

2. Then, the topic distribution $P(z|d)$ is fixed, and PLSA is run on the visual words to compute $P(v|z)$. Again, the EM algorithm is used for optimization:

$$\mathcal{L}_V = \prod_{d \in D} \left[P(d) \cdot \prod_{v} P(v|d)^{n(v,d)} \right]. \tag{4}$$

3.1.2 *Inference*

Given a previously unseen batch D^* of test images d^* to be labeled, the baseline approach treats all images independently. Thereby, the visual words of image d^* are given, and the tag distribution $P(t|d^*)$ is inferred by inverting the training procedure: first – given $P(v|d^*)$ – $P(z|d^*)$ is computed using the EM algorithm for $\mathcal{L}_V$, whereas the topics $P(v|z)$ learned in training are kept fixed. Then the distribution of tags is estimated as:

$$P(t|d^*) = \sum_{z} P(t|z) \cdot P(z|d^*) \tag{5}$$

A set of tags with maximum posterior probability is finally selected as annotations of d^*.

3.2 Style Variant 1: Appearance Only

In this section, we present an extension of the baseline approach from Section 3.1 for labelling picture batches that share a coherent style. Thereby, style is modeled as a latent random variable $s \in S$ (as discussed previously, these styles might correspond to different holiday types like "sightseeing" or "nature trip"). Images of a

batch are assumed to be independent samples drawn from the same style-specific model. We assume that style can be inferred reliably from the whole batch, and that a style-specific model gives a more accurate image annotation.

The visual word distribution $P(v|z)$ from the baseline model is replaced with style-specific appearance models $P(v|z,s)$. Then, the visual word distribution of an image d with style s can be rewritten as:

$$P(v|d,s) = \sum_z P(v|z,s) \cdot P(z|d) \tag{6}$$

The tag distribution $P(t|z)$ remains unchanged, i.e. we assume that the frequency with which a tag appears does not depend on the style. Appearance, however, may differ between styles (for example, a "building" in an "Africa" style and in a "New York City" style may look different). We will refer to this model as the "appearance only" style model in the following.

3.2.1 Learning

Like for the baseline model, a two-step learning procedure is used similar to the one in Section 3.1.1. First, standard EM on all input images (regardless of style) is used to learn $P(t|z)$ and $P(z|d)$ by maximizing the tag likelihood (Equation (4)). Second, the distribution of visual words $P(v|z,s)$ is learned separately for each style. For this purpose, we assume that the style $s(d)$ for each training image d is given. For each style s, the following likelihood is maximized:

$$\mathcal{L}_V^s = \prod_{d:s(d)=s} \left[P(d) \cdot \prod_v P(v|d,s)^{n(v,d)} \right]. \tag{7}$$

Again, optimization is carried out using EM, whereas the topic distributions $P(z|d)$ are kept fixed.

3.2.2 Inference

Compared to inference in the baseline model (Section 3.1.2), the key difference is that the style variable s is unknown. As Sarkar and Nagy demonstrate, globally optimal Bayesian inference of tags and style is usually infeasible [23]: since tags and style are both unknown and influence each other, optimal inference requires to test all combinations of tags, whose number grows exponentially with the number of test images in a batch. To resolve this problem, we follows a similar strategy as in [23]: it is assumed that – for image batches of sufficient size – the style parameter can be reliably inferred by maximizing the visual word likelihood:

$$s^* = \arg\max_s \left[\prod_{d^* \in D^*} \left(P(d^*) \cdot \prod_{v \in V} P(v|d^*,s)^{n(v,d^*)} \right) \right] \tag{8}$$

This leads to an annotation procedure in which the appearance likelihood (Equation (8)) is computed for each style, and after this a style-specific annotation is run for the best style s^*.

3.3 Style Variant 2: Appearance and Tags

The appearance-only style model from Section 3.2 makes limited use of style information in a sense that the distribution of tags is still style-independent. In practice, however, tags may be strongly correlated with style (for example, the tags given to pictures from a New York City trip may differ significantly to the ones given to pictures from a visit to Rome). To exploit this information to its full potential, a second style variant is proposed in which both appearance and tags are modeled by style-dependent distributions $P(v|d,s)$ and $P(t|d,s)$. This leads to a set of entirely decoupled

Figure 3: A random sample of pictures from our FLICKR dataset, which consists of 8,000 images from 8 Flickr groups like "Africa" or "New York City".

style-specific annotation models, i.e. each style is trained and applied independently.

3.3.1 Learning and Inference

Since styles are completely decoupled, training simplifies to learning a separate PLSA-based annotation model per style. Similar to the baseline from Section 3.1.1 the EM algorithm is used, only that the distributions $P(v|d)$ and $P(t|d)$ are replaced with style-specific equivalents $P(v|d,s)$ and $P(t|d,s)$ that are trained on style-specific training image sets $\{d|s(d)=s\}$.

For inference, annotation is run for all styles, and the target style is determined using the same maximum likelihood criterion as for the appearance-only style model in Equation (8) (only that $P(t|d)$ is replaced by its style-specific equivalent $P(t|d,s)$).

4. EXPERIMENTS

In this section, we evaluate the proposed combination of style modeling and image annotation in quantitative experiments. The use case scenario is a tagging of holiday snapshots – different styles of holidays (like "New York" or "African Safari") are learned from Flickr groups and then used to annotate style-consistent batches of personal pictures.

We first present results on images from the COREL dataset and from the photo-sharing website Flickr, in which we compare our approach with several canonical baselines and oracle-based control runs. Afterwards, our method is compared to results from the literature on the COREL-5K image annotation benchmark [7]. The grouping information required by our style modeling is provided in form of COREL folders or FLICKR groups.

4.1 Experiment 1: COREL and FLICKR

The goal of this experiment is to provide an in-depth analysis of style modeling and a comparison with several canonical baselines. We first describe the experimental setup, starting with the datasets:

Small size COREL (COREL-13) 13 folders (1,300 images overall) were selected from the COREL dataset corresponding to countries, regions, and cities (for example, "Africa" and "Kyoto"). Images in the same folder are assigned to the same batch, i.e. assumed to share the same style. A vocabulary of 644 tags from the COREL annotations was used. Since the average number of tags per image is 4, we select the 4 tags with the highest scores $P(t|d^*)$ as annotations.

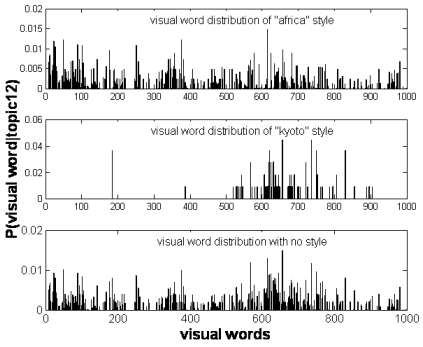


Figure 4: The visual word distribution of Topic No. 12 for the styles "Africa" (top), "Kyoto" (center), and without modeling style ("bottom"). It can be seen that style has a massive influence on the appearance of a topic.

Large size COREL (COREL-45) To evaluate the performance of style modeling for more styles, a dataset similar to COREL-13 was sampled, only that 45 folders (4,500 images overall) were used. The tag vocabulary size is 1,257. Similar to COREL-13, the top 4 tags are selected as annotations.

FLICKR This dataset contains 8,000 images downloaded from Flickr. Style corresponds to *Flickr groups* [21] representing travel scenarios like "New York Sightseeing" and "African Safari" (for sample pictures, please refer to Figure 3). 8 styles were used with 1,000 images each. A vocabulary of 544 tags was created from the most frequent original Flickr tags by filtering unsuitable tags (like "d40", "2008", or "Olympus"). Since the average number of tags per image is about 6, the 6 most probable tags are returned as annotations by each method.

For all three datasets, visual words were sampled from the images using a dense regular sampling of SIFT features at several scales, giving ca. 4,800 patches per image on average. These were clustered to 2,000 visual words using K-Means (we used a fast version [8] available from `http://mloss.org`). Test results were averaged over multiple runs (20 for COREL-13 and FLICKR, 5 for COREL-45), whereas in each run a random stratified split into 80% training and 20% testing was done. Test images were joined to style-consistent batches of 20 pictures each. As a measure of annotation performance, the F-measure (weighted harmonic mean of precision and recall) is used. For each test image d^*, the annotation result $T_R(d^*)$ is compared with the ground truth annotations $T_{GT}(d^*)$, computing the image-wise precision $P(d^*)$ and recall $R(d^*)$:

$$P(d^*) = \frac{|T_R(d^*) \cap T_{GT}(d^*)|}{|T_R(d^*)|}, \quad R(d^*) = \frac{|T_R(d^*) \cap T_{GT}(d^*)|}{|T_{GT}(d^*)|}$$

By averaging these over all test images, the *mean* image-wise precision $\bar{P}$ and recall $\bar{R}$ are obtained. These are finally combined to the F-measure:

$$\text{F-measure} = 2 \cdot (\bar{P} \cdot \bar{R})/(\bar{P} + \bar{R})$$

Apart from the final annotation results, we also evaluate the performance of the maximum likelihood style decision by displaying the

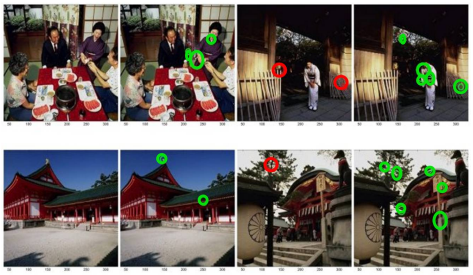


Figure 6: The visual words corresponding to Topic No. 12, which is strongly linked to the tags "people" and "temple". The non-style result is on the left (red patches), the "Kyoto" style on the right (green). For the style model, more patches can be found that activate the topic, and better tagging results can be expected.

accuracy P_{style} over varying image batch size:

$$P_{style} = \frac{\#image\ batches\ assigned\ correct\ style}{\#image\ batches}$$

Eight different methods were tested, whereas the number of topics was fixed to $|Z| = 20$ (which gave the best results in previous tests):

Baseline, few topics This is the plain PLSA annotation model from Section 3.1 using $|Z|$ topics. Images are tagged independently, and style information is discarded.

Baseline, many topics To make sure that potential performance improvements with style are not attributed to a higher number of topics, we also test the baseline with $|Z| \cdot |S|$ topics (which equals the overall number of topics in the appearance-and-tags style model).

Appearance-only style, by single image The model from Section 3.2. The batch size is set to 1, i.e. each image is mapped to a style and labeled independently. Serves as another baseline.

Appearance-only style, by batch The same model, but now style is decided based on the whole batch.

Appearance-only style, by assignment The same model, but the correct style is assigned automatically (serves as an oracle-based control experiment to estimate performance loss due to incorrect style assignment).

Appearance and tag style, by single image The model from Section 3.3. Style is assigned for each image individually (serves as a baseline).

Appearance and tag style, by batch The same model, but now style is decided based on the whole batch.

Appearance and tag style, by assignment The same model, but the correct style is assigned automatically (serves as a control experiment).

We start the results with a short illustration of style modeling by validating a fundamental assumption, namely that style-specific topic models differ from their non-style equivalents. This is demonstrated for a sample topic and the two styles "Africa" and "Kyoto" from the COREL dataset. We pick a topic (referred to as "Topic No.

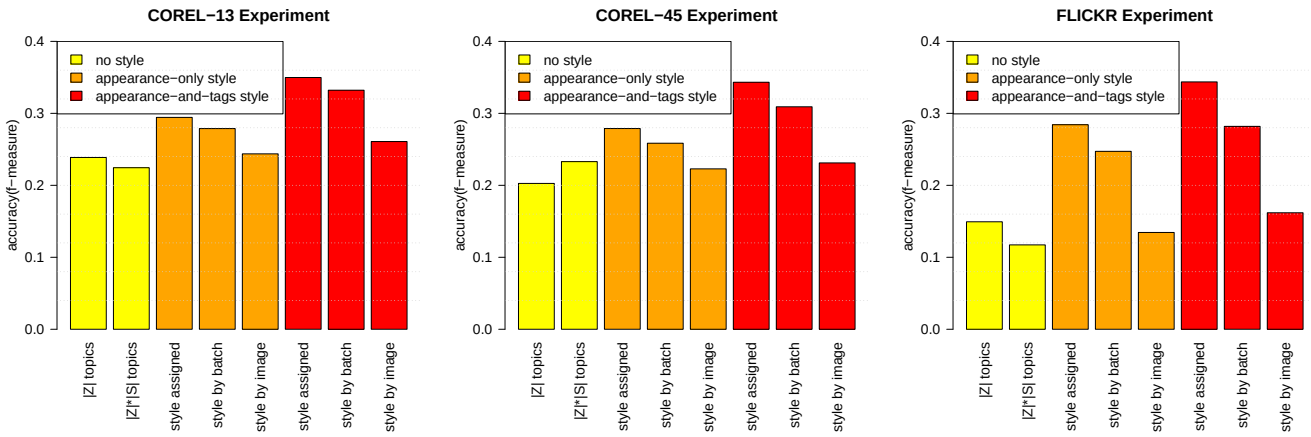


Figure 5: Quantitative experimental results for Experiment 1. Key observations are that style modeling improves performance significantly, that appearance-and-tag style modeling performs best, and that tagging based on the batch always outperforms single-image tagging.

12" in the following) whose most frequent tags contain the words "people" and "temple". In Figure 4, we visualize the distribution of visual words $P(v|''topic\ 12'')$ for the baseline model, as well as $P(v|''topic\ 12'', ''africa'')$ and $P(v|''topic\ 12'', ''kyoto'')$ (for the appearance-only style approach). Obviously, the appearance learned for the two styles differs strongly from the one in the global model.

The next question is how style modeling influences the fitting between topics and image appearance. This is illustrated in Figure 6, which shows sample images related to the tags "people" and "temple" associated with Topic No. 12. The visual words with highest topic scores $P(v|''topic\ 12'') \geq 0.5$ are highlighted for the non-style case (left, red) and the style "Kyoto" (right, green). While in the baseline model only few patches can be found related to the topic, for the style approach multiple patches activate Topic No. 12, and a better annotation result can be expected.

Quantitative results for all three datasets are given in Figure 5. These plots provide several insights: first, style modeling improves annotation performance significantly compared to the non-style baselines: For both versions of the baseline, relative improvements between 46.5% (COREL-13) and 130.1% (FLICKR) are measured. Second, appearance and tags style outperforms the appearance-only style model on all datasets. Again, these relative improvements are significant, ranging from 14.0% (FLICKR) to 19.6% (COREL-45). Third, some performance loss occurs due to incorrect style decisions. This can be observed when comparing the "style by batch" versions with the oracle-based "style by assignment" control runs. For all datasets and style versions, a moderate performance loss can be observed ranging from 5.3% (COREL-13, appearance-and-tag style) to 12.9% (FLICKR, appearance-and-tag style). When comparing the COREL-13 and COREL-45 runs, it can be seen that this performance loss increases slightly with the number of styles. This can be attributed to an increasing confusion due to a higher number of styles (correspondingly, the accuracy of style decision decreases from 83.1% to 65.8%). Overall, the benefits of style modeling in terms of annotation performance decrease slightly when scaling from 13 to 45 styles, but remain significant. Finally and most importantly, batch-wise annotation outperforms image-wise annotation in all cases, i.e. it helps to use the style consistency of an image batch.

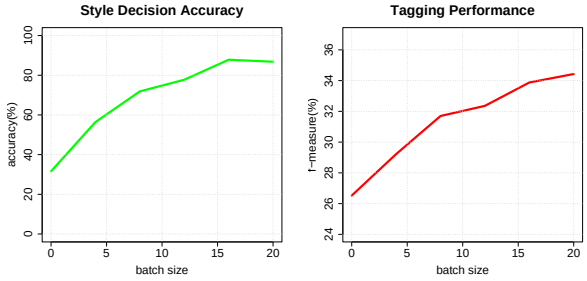


Figure 7: Both the accuracy of style decision (left) and the overall annotation performance (right) increase with the test batches' size. The leftmost point in both plots corresponds to a tagging of individual images.

This can be seen when comparing the "style-by-batch" results with the "style-by-image" ones. Relative performance improvements range from 14.4% (COREL-13, appearance-only style) to 83.8% (FLICKR, appearance-only style).

It seems that annotation performance is correlated with test batches' size, as the *style decision* becomes more reliable with increasing image batch size. This fact is supported by Figure 7, which plots both the accuracy of style decision and the overall annotation performance for the COREL-13 test (using the tag-and-appearance style model and averaging over 10 cross-validation runs). Image batches of varying size are used in testing. It can be seen that – by increasing batch size from 1 to 20 – the style decision accuracy can be improved significantly from 31.6% to 86.8%, and correspondingly the annotation performance (F-measure) increases from 26.5% for 34.4%. Even for a rather small batch size of 8 images, a relative performance improvement of 20% is achieved.

Finally, we tackle the question what styles tend to be confused most often. Figure 8 illustrates the confusion matrix of style decision on the COREL-45 dataset. An in-depth inspection reveals that the most frequently confused classes do in fact show an intuitive resemblance: for example, style 18 ("Kenya") and style 2

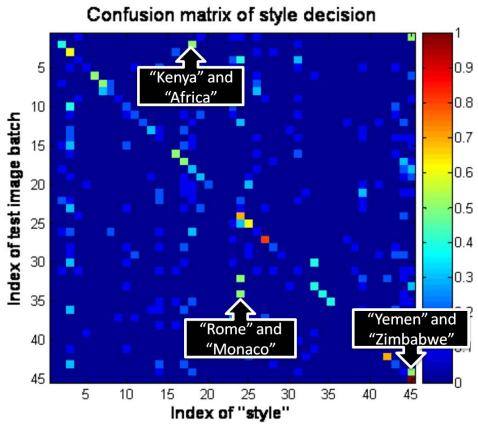


Figure 8: The confusion matrix of style decision on the COREL-45 dataset. The most frequent confusions tend to occur for visually similar styles (like "Kenya" vs. "Africa")

("Africa") are often confused by our system (the probability is near 0.5). Other frequent confusions are style 34 ("Rome") with style 24 ("Monaco"), both showing similar buildings, or style 44 ("Yemen") with style 45 ("Zimbabwe").

5. EXPERIMENT 2: COREL-5K

In a second experiment, the proposed style consistency model is compared to other approaches from the literature. Therefore, the popular COREL-5K benchmark for image annotation is used [4, 7, 10, 14, 25]. The dataset consists of 5,000 images from the COREL dataset corresponding to 50 folders of 100 images each. Like in the previous COREL tests, a 1:1 correspondence between styles and folders is imposed. The dataset is split into a training set of 4500 images (90 images per style) and a test set of 500 images (10 images per style), i.e. the batch size is set to 10. Similar to previous tests in the literature, we use the same tag vocabulary of 374 terms and return the top 5 words as annotation results. The original COREL images were downscaled to a width of 192 pixels (we want to thank R. Manmatha and Shaolei Feng for providing the dataset). Visual words were extracted by a regular sampling of ca. 5,400 patches of side length 12 per image. These were described using DCT coefficients in YUV space. A K-Means clustering to 2,000 visual words was used.

A few sample annotations of the proposed style model in Figure 9 illustrate the benefits of style modeling for image annotation. As quantitative results, we report the same performance measures as used in the literature: for each tag t, the per-word precision and per-word recall are measured over all test images d^*. These values are averaged over all 251 tags occurring in the test set to obtain the mean per-word precision and recall. Further, the number of tags t with $R(t) > 0$ is reported. The results in Table 1 include a variety of figures reported by other researchers: the co-occurrence model by Mori et al. [19], the machine translation model from Duygulu et al. [7], two relevance models by Manmatha and co-workers [10, 14], supervised multi-class labeling by Carneiro et al. [4], and several other annotation models [25, 5]. Our tests also include two baseline approaches run by ourselves: the non-style PLSA model [18] (Section 3.1) – which does not employ style consistency – and the appearance-and-tags style model applied to images individually. Both baselines show a low performance (F-measures 5% / 16%). However, by tagging images in style-consistent batches, the proposed approach achieves the best result reported on the COREL-5k benchmark so far, with a recall of 39%, precision of 25%, and F-measure of 31%. It should be noted, though, that this is achieved by using the structure of image content, a source of information which is neglected by all other approaches in Table 1. Note also that this improvement cannot be attributed to the underlying annotation model (which by itself performs rather poorly), but is clearly due to the exploitation of style consistency. Consequently, it can be expected that other probabilistic annotation models (like [10] or [4]) could benefit from style consistency modeling in a similar fashion.

Table 1: A comparison with methods from the literature on the COREL-5k benchmark. By making use of style consistency, the proposed approach achieves the best result reported on the benchmark so far.

Approach	#words with rec.>0	mean prec.	mean rec.	F-measure
co-occurrence [19, 4]	19	0.02	0.03	0.02
Translation [7, 4]	49	0.04	0.06	0.05
kernel densities with tag co-occurrence [5]	91	0.11	0.13	0.12
SVDCos [25]	102	0.15	0.15	0.15
CRM [14]	107	0.16	0.19	0.17
CSD-Prop [25]	130	0.20	0.27	0.23
MBRM [10]	122	0.24	0.25	0.24
SML [4]	137	0.23	0.29	0.26
CSD-SVM [25]	127	**0.25**	0.28	0.26
PLSA (no style) [18]	57	0.04	0.09	0.05
PLSA (by image)	106	0.13	0.23	0.16
PLSA (by batch)	**141**	**0.25**	**0.39**	**0.31**

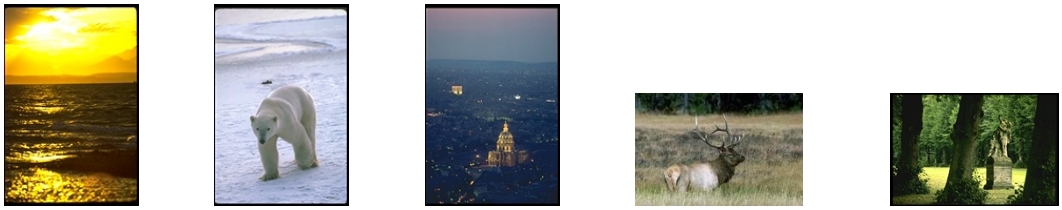
6. DISCUSSION

In this paper, we have presented an image annotation framework making use of the fact that pictures to be labeled often come in groups of coherent appearance. An approach from the domain of handwriting recognition – which can generally be used as a wrapper around probabilistic image annotation approaches – was demonstrated to improve annotation results significantly compared to tagging images individually. Alternatively, our method could be used as a style inference approach to provide users with group recommendations for their pictures.

Potential future directions along the proposed line of research are the integration with other style models to overcome the need for discrete styles [17], or a combination with other approaches that may in a broader variety of practical use cases provide the grouping information that was assumed to be given in this work. One information source to infer such a grouping is meta-data such as capture time and location [3]. Alternatively, it might be interesting to learn styles using a content-driven clustering [2]. Both approaches come with the benefit of a fully automatic learning free of external style information.

7. REFERENCES

[1] H. Baird and G. Nagy. A Self-Correcting 100-Font Classifier. *Proc. SPIE — The International Society for Optical Engineering*, 2181:106–115, 1994.



sun window courtyard palace church	**snow** clouds jet plane fox	water boats autumn oahu maui	**grass** house village castle zebra	bridge flower market petals nest
sun water waves tree plane	**bear polar snow ice** head	**city sky buildings** mountain sunset	**grass elk deer moose** white-tailed	**statue palace garden** people sunset

Figure 9: Sample annotation results on the COREL-5K benchmark, comparing annotation results without style (top) and with style (bottom). It can be seen that style modeling improves tagging performance significantly: for example, the fourth image belongs to a batch that is correctly assigned to a "nordic wildlife" style, such that the correct tags "elk" and "moose" can be inferred.

[2] K. Barnard and D. Forsyth. Learning the Semantics of Words and Pictures. In *Proc. Int. Conf. Computer Vision*, pages 408–415, July 2001.

[3] L. Cao, J. Luo, H. Kautz, and T. Huang. Annotating Collections of Photos using Hierarchical Event and Scene Models. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[4] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[5] A. L. Coto and S. Rüger. Can a Probabilistic Image Annotation System be Improved using a Co-occurrence Approach? In *Proc. SAMT Workshop on Cross-Media Information Analysis and Retrieval*, December 2008.

[6] M. Cristani, A. Perina, U. Castellani, and V. Murino. Geo-Located Image Analysis using Latent Representations. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, June 2008.

[7] 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. Europ. Conf. Computer Vision*, pages 97–112, May 2002.

[8] C. Elkan. Using the Triangle Inequality to Accelerate KMeans. In *Proc. Int. Conf. Machine Learning*, pages 147–153, August 2003.

[9] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Technical report, PASCAL Challenge Workshop, October 2007.

[10] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 1002–1009, June 2004.

[11] A. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen. Image Annotation using Personal Calendars as Context. In *Proc. Int. Conf. Multimedia*, pages 681–684, October 2008.

[12] T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196, 2001.

[13] H. Kück, P. Carbonetto, and N. de Freitas. A Constrained Semi-Supervised Learning Approach to Data Association. In *Proc. Europ. Conf. Computer Vision*, pages 1–12, May 2004.

[14] V. Lavrenko, R. Manmatha, and J. Jeon. A Model for Learning the Semantics of Pictures. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[15] J. Leon, V. Lavrenko, and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-media Relevance Models. In *Proc. Int. SIGIR Conf. Research and Development in Information Retrieval*, pages 119–126, July 2003.

[16] J. Li and J. Wang. Real-time Computerized Annotation of Pictures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[17] C. Mathis and T. Breuel. Classification Using a Hierarchical Bayesian Approach. In *Int. Conf. Pattern Recognition*, pages 103–106, August 2002.

[18] F. Monay and D. Gatica-Perez. PLSA-based Image Annotation: Constraining the Latent Space. In *Proc. Int. Conf. on Multimedia*, pages 348–351, October 2004.

[19] Y. Mori, T. Takahashi, and R. Oka. Image-to-Word Transformation based on Dividing and Vector Quantizing Images with Words. In *Proc. Int. Workshop Multimedia Intelligent Storage and Retrieval Management*, 1999.

[20] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging Context to Resolve Identity in Photo Albums. In *Proc. Joint Conf. Digital Libraries*, pages 178–187, June 2005.

[21] R. Negoescu and D. Gatica-Perez. Analyzing Flickr Groups. In *Proc. Int. Conf. Image and Video Retrieval*, pages 417–426, July 2008.

[22] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A Thousand Words in a Scene. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

[23] P. Sarkar and G. Nagy. Style Consistent Classification of Isogenous Patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(1):88–98, 2005.

[24] J. Sivic and A. Zisserman. *Video Google: Efficient Visual Search of Videos*, pages 127–144. Springer, 2006.

[25] J. Tang and P. Lewis. A Study of Quality Issues for Image Auto-Annotation With the Corel Dataset. *IEEE Trans. Circuits Syst. Video Techn.*, 17(3):384–389, 2007.

[26] C. Yang, M. Dong, and F. Fotouhi. Region Based Image Annotation through Multiple-instance Learning. In *Proc. Int. Conf. Multimedia*, pages 435–438, November 2005.