Multimed Tools Appl
DOI 10.1007/s11042-012-1096-y

With a little help from my friends

Community-based assisted organization of personal photographs

Claudio Cusano · Simone Santini

Abstract In this paper, we propose a content-based method the for semi-automatic organization of photo albums based on the analysis of how different users organize their own pictures. The goal is to help the user in dividing his pictures into groups characterized by a similar semantic content. The method is semi-automatic: the user starts to assign labels to the pictures and unlabeled pictures are tagged with proposed labels. The user can accept the recommendation or made a correction. To formulate the suggestions is exploited the knowledge encoded in how other users have partitioned their images. The method is conceptually articulated in two parts. First, we use a suitable feature representation of the images to model the different classes that the users have collected, second, we look for correspondences between the criteria used by the different users. Boosting is used to integrate the information provided by the analysis of multiple users. A quantitative evaluation of the proposed approach is obtained by simulating the amount of user interaction needed to annotate the albums of a set of members of the flickr® photo-sharing community.

Keywords Personal photography · Automatic image annotation · Content-based image analysis · Social image retrieval

S. Santini was supported in part by the *Ministerio de Educación y Ciencia* under the grant N. MEC TIN2008-06566-C04-02, *Information Retrieval on different media based on multidimensional models: relevance, novelty, personalization and context*.

C. Cusano (✉)
Department of Informatics, Systems and Communication (DISCo),
Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
e-mail: claudio.cusano@disco.unimib.it

S. Santini
Escuela Politécnica Superior, Universidad Autónoma de Madrid,
C/ Tomas y Valiente 11, 28049 Madrid, Spain
e-mail: simone.santini@uam.es

Published online: 04 May 2012

1 Introduction

One of the most remarkable social and technical by-products of the diffusion of the internet is the emergence of communities connected not by physical proximity but by common interests. While these utopic (in the original etymology of $ου - τοπος$: no place) communities have existed for a long time, from the medieval monastic orders to the scientific community, the internet has given them a stronger cohesion by providing the technical instruments for frequent communication. Socially, one fairly evident consequence of the internet has been, in a sense, the trivialization of the common interest around which communities gather. In other times, the survival of an utopic community required a considerable effort, and could be justified only by a common interest that its participants regarded as primary. With the advent of the internet, communities are created easily, and every one of us can be at the same time a member of many of them, often representing interests that we regard as superficial and of little importance. Technically—this is the aspect of interest here—these communities have created a great interest in peer-to-peer systems and in *social filtering*, a technical instrument to use the collective wisdom, so to speak, of the community to the advantage of each one of its members.

This paper will consider a community that is not held by very strong ties but to which many of us, at some time or another, belong: that of amateur photographers. We will consider a community of people interested in non-commercial photography who place their photographs in a suitable web server (flickr®, or similar services) where they can be shared by a community of similarly interested people. One of the most common activities in which people engage when organizing pictures is that of classification: the pictures in the camera will be divided into thematic groups. The criteria that preside this organization are highly personal: in this case, what's good for the goose is not necessarily good for the gander. The same vacation photos that a person will divide in "Rhodos" and "Santorini" will be divided by someone else into "family", "other people" and "places" or into "beach", "hotel" and "excursion", or in any other organization. Our purpose is to use the "collective wisdom" consituted by all these categorizations to help a new user classify his pictures. Consider a new user (which we call the *apprentice*) who is trying to classify vacation pictures. In a folder, she will start placing pictures that, visually, have little consistence: there will be some photos of a beach, some close and medium shots of family members, some pictures taken in an hotel or a campsite. In a different folder, the same apprentice wants to place images of a visit to Rome. Suppose that there are other users (yclept the *wizards*) who already have created categories, and who have several images in each category. One of them (call her "wizard A") has, among others, a category of beach images, while the other (call her "wizard B") has a category corresponding to a trip to Rome. As soon as the apprentice will start categorizing images, the system will realize that one of the categories contains images similar to the vacation folder of wizard A. Wizard A will then be used as a classifier to suggest new pictures that can be placed in the vacation folder of the apprentice. Similarly, wizard B will be used as a classifier to suggest pictures for the "Rome" category.

We can see a system like this under two possible lights. On one hand, we can see it as a classification aid. In this view, the apprentice has a certain classification in mind, which she will not change, and the purpose of the system is to help her by bringing up-front, in a suitable interface, the pictures that will go into the folders that

the apprentice has created. On the other hand, we can see it as an exploration and discovery tool. When the apprentice begins making the classification, her ideas are still uncertain, and she will be open to changes and adaptations of her scheme. In this sense, bringing up photos according to the classification scheme of the wizards will create a dialectic process in which criteria are invented, discarded, modified. The classification with which the apprentice will end up with mightn't remind the original one at all, simply because looking at the organization induced by the wizards has given her new ideas.

This second view is, in many ways, the most interesting one. Alas, it is virtually impossible to evaluate the effectiveness of a system in this capacity short of long term user satisfaction studies. As a matter of praxis, in this paper we will only consider our system in the first capacity: as an aid to create a fixed classification, and will evaluate it accordingly.

1.1 Related work

A number of commercial products is available for the management and organization of personal photo collections. In spite of being convenient and user friendly, these products still rely largely on manual annotation for browsing and retrieval. To overcome this limitation several automatic or semi-automatic content-based approaches have been proposed. A prototype system for home photo management and processing has been implemented by Sun et al. [17]. Together with traditional tools, they included a function to automatically group photos by time, visual similarity, image class (indoor, outdoor, city, landscape), or number of faces (as identified by a suitable detector).

1.1.1 Annotation and meta-data

Another system for managing family photos has been developed by Wenyin et al. [22]. The system allows the categorization of photos into some predefined classes. A semi-automatic annotation tool, based on retrieval by similarity, is also provided. When the user imports some new images, the system searches for visually similar archived images. The keywords with higher frequencies in these images are used to annotate the new images. Keywords have to be confirmed or rejected in a successive retrieval-feedback process. Mulhem and Lim proposed the use of temporal events for organizing and representing home photos using structured document formalism [12]. Retrieval and browsing of photos are based on both temporal context and image content, represented by the occurrence of 26 classes of visual keywords. Shevade and Sundaram presented an annotation paradigm that attempts to propagate semantic by using WordNet and low-level features extracted from the images [16]. As the user begins to annotate images, the system creates positive and negative example sets for the associated WordNet meanings. These are then propagated to the entire database, using low-level features and WordNet distances. The system then determines the image that is least likely to have been annotated correctly and presents the image to the user for relevance feedback.

A common approach for automatic organization of photo albums consists in the application of clustering techniques to group images into visually similar sets. Manual post-processing is usually required to modify the clusters in order to match

user's categorization. Information about time is often used to improve clustering by segmenting the album into events. Platt proposed a method for clustering personal images taking into account timing and visual information [15]. Loui and Savakis described an event-clustering algorithm which automatically segments pictures into events and sub-events, based on date/time metadata information, as well as color content of the pictures [10]. Li et al. exploited time stamps and image content to partition related images in photo albums [8]. Key photos are selected to represent a partition based on content analysis and then collated to generate a summary. A semi-automatic technique has been presented by Jaimes et al. [7]. They used the concept of Recurrent Visual Semantics (the repetitive appearance of visually similar elements) as the basic organizing principle. They proposed a sequence-weighted clustering technique which is used to provide the user with a hierarchical organization of the contents of individual rolls of film. As a last step, the user interactively modifies the clusters to create digital albums.

1.1.2 Faces

Since people identity is often the most relevant information for the user, it is not surprising that several approaches have been proposed for the annotation of faces in family albums. Das and Loui used age/gender classification and face similarity to provide the user with the option of selecting image groups based on the people present in them [3].

Another framework for semi-automatic face annotation has been proposed by Chen et al. [2]. In addition to the traditional face recognition features they used similarity search and relevance feedback on a set of color and texture features. Zhang et al. have reformulated the face annotation from a pure recognition problem to a problem of similar face search and annotation propagation [23]. Their solution integrates content-based image retrieval and face recognition algorithms in a Bayesian framework.

1.1.3 Narrative and community

More recently, some systems have begun to consider photo organization from the point of view of evaluating a whole set of images, rather than classifying individual images. For example, in [14], groups of images are created with an eye on the storytelling quality of the whole group rather than the fitness of individual images.

The idea of exploiting user correlation in photo sharing communities has been investigated by Li et al. [9]. They proposed a method for inferring the relevance of user-defined tags by exploiting the idea that if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of visual content. Each tag of an image accumulates its relevance score by receiving votes by neighbors (i.e. visually similar images) labeled with the same tag.

2 Method

In this paper, we propose a method the for semi-automatic organization of photo albums. The method is content-based, that is, only pictorial information is considered. It should be clear from the contents of the paper that the method is applicable to

non-visual information such as keywords and annotations. In spite of the importance that these annotations may have for the determination of the semantics of images, we have decided to limit our considerations to visual information on methodological grounds, since this will give us a more immediate way of assessing the merits of the method vis-à-vis simple similarity search.

The goal is to help the user in classifying pictures dividing them into groups characterized by similar semantics. The number and the definition of these groups are completely left to the user.

This problem can be seen as an on-line classification task, where the classes are not specified a priori, but are defined by the user himself. At the beginning all pictures are unlabeled, and the user starts to assign labels to them. After each assignment, the unlabeled pictures are tagged with proposed labels. The user can accept the recommendation or make a correction. In either case the correct label is assigned to the image and the proposed labels are recomputed. Unlabeled pictures are displayed sorted by decreasing confidence on the correctness of the suggestion, but the order in which the user processes the images is not restricted. Provided a reasonable user interface is available, the labels proposed by the method can be confirmed very quickly, allowing for a rapid and convenient organization of the album. Figure 1 shows two screenshots of a prototypical system which implements the proposed method.

2.1 Exploiting users' correlation

One of the difficulties of assisted album organization is that, at the beginning, we don't have any information on the criteria that the user is going to apply in

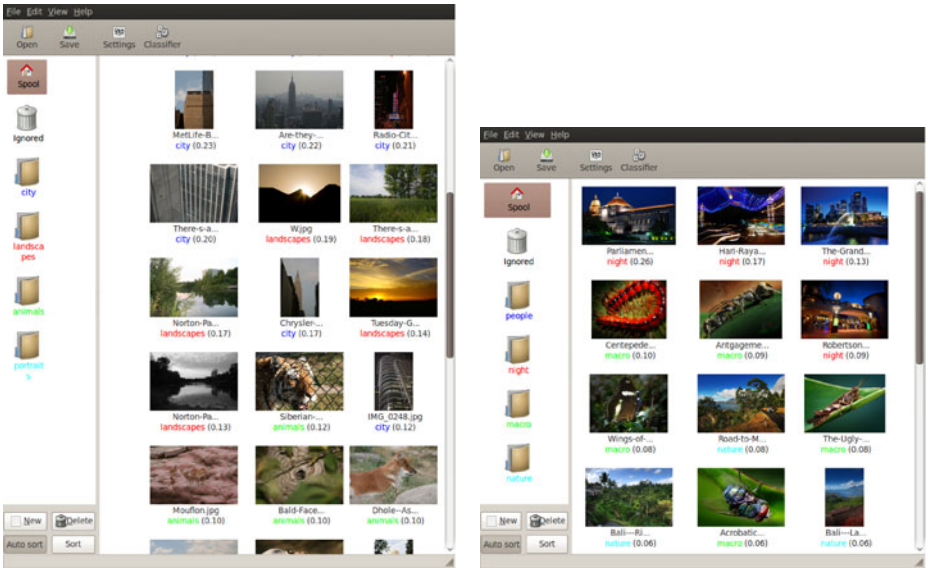

Fig. 1 Screenshots of the image annotation tool. Each unlabeled picture is annotated with a class proposed by the system. Proposals are chosen among the classes defined by the user (represented by the folders on the left side). The confidence scores about the proposals are used to sort the images

partitioning his pictures. However, a huge library of possible criteria is available in photo-sharing communities. The users of these services are allowed to group their own images into sets and we can assume that these sets contain pictures with some characteristic in common. For instance, sets may contain pictures taken in the same location, or portraying a similar subject.

Our idea is to exploit the knowledge encoded in how a group of users (*wizards*, in the following) have partitioned their images, in order to help organize the pictures of a different user (the *apprentice*). The method is conceptually articulated in two parts. First, we use a suitable feature representation of the images of the wizards to model the different classes that they have collected, second, we look for correspondences between the (visual) criteria used in the wizards' classes and those that the apprentice is creating in order to provide advice. In other, somewhat oversimplistic words; if we notice that one of the classes that the apprentice is creating appears to be organized using criteria similar to those used in one or more wizard's classes, we use the wizards' classes as representative, and the unlabeled apprentice images that are similar to those of the wizard class are given the label of that class.

Consider a wizard, who partitioned his pictures into the C categories $\{\omega_1, \ldots, \omega_C\} = \Omega$. These labeled pictures are used as a training set to train a classifier that consists of a classification function $g : X \to \Omega$ from the feature space X into the set of user defined classes. If the partition of the wizard exhibits regularities (in terms of visual content) that may be exploited by the classification framework, then g may be used to characterize the pictures of the apprentice as well. Of course, it is possible that the apprentice would like to organize his pictures into different categories. However, people tend to be predictable, and it is not at all uncommon that the sets defined by two different users present some correlation that can be exploited. To do so, we define a mapping $\pi : \Omega \to Y$ between the classes defined by the wizard and the apprentice (where $Y = \{y_1, \ldots, y_k\}$ denotes the set of apprentice's labels). We allow a non-uniform relevance of the apprentice's images in defining the correlation with the wizard's classes. Such a relevance can be specified by a function w that assigns a positive weight to the images. Weighting will play an important role in the integration of the predictions based on different wizards, as described in Section 2.3. Let $Q(\omega_i, y_j)$ be the set of images to which the apprentice has assigned the label y_j, and that, according to g, belong to ω_i; then π is defined as follows:

$$\pi(\omega) = \arg\max_{y \in Y} \sum_{x \in Q(\omega, y)} w(x), \quad \omega \in \Omega, \tag{1}$$

where a label is arbitrarily chosen when the same maximum is obtained for more than one class. That is, π maps a class ω of the wizard into the class of the apprentice that maximizes the cumulative weight of the images that g maps back into ω. If no apprentice image belong ω we define $\pi(\omega)$ to be the class of maximal total weight.

If we interpret w has a misclassification cost, our definition of π denotes the mapping which, when combined with g, minimizes the total misclassification error on the images of the apprentice:

$$\min_{\pi : \Omega \to Y} \sum_{x,y} w(x) \left(1 - \chi_{\{y\}}(\pi(g(x)))\right), \tag{2}$$

where the summation is taken over the pairs (x, y) of images of the apprentice with the corresponding labels, and where χ denotes the indicator function ($\chi_A(x) = 1$ if $x \in A$, 0 otherwise). Figure 2 depicts and summarizes how π is defined.

The composition $h = \pi \circ g$ directly classifies elements of X into Y. In addition to embedding the correlation between the wizard and the apprentice, the design of h enjoys a useful property: the part defined by g is independent of the apprentice, so that it can be computed off-line allowing for the adoption of complex (and hopefully accurate) machine learning models such as SVMs, neural networks, and the like; the part defined by π, instead, can be worked out very quickly since its computation is linear in the number of the images labeled by the apprentice and does not depend on the whole album of the wizard, but only on its partial representation provided by g.

In this work, g is a $k-$nearest neighbor (KNN) classifier. Other classification techniques may be used as well, and some of them would probably lead to better results. We decided to use the KNN algorithm because it is simple enough to let us concentrate on the correlation between the users, which is the main focus of this paper.

2.2 Image description

Since we do not know the classes that the users will define, we selected a set of four features that give a fairly general description of the images. We considered two features that describe color distribution, and two that are related to shape information. One color and one shape feature are based on the subdivision of the images into sub-blocks; the other two are global. The four selected features are: spatial color moments, color histogram, edge direction histogram, and a bag of features histogram.

Spatial color distribution is one of the most widely used feature in image content analysis and categorization. In fact, some classes of images may be characterized in terms of layout of color regions, such as blue sky on top or green grass on bottom. Similarly to Vailaya et al. [19], we divided each image into 7×7 blocks and computed

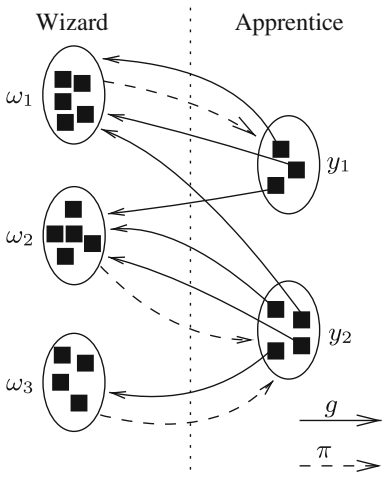

Fig. 2 Example of definition of the mapping π between a wizard and the apprentice, assuming uniform weights. Since g maps into ω_1 two images of class y_1 and only one image of class y_2, we have that $\pi(\omega_1) = y_1$. Similarly, $\pi(\omega_2) = \pi(\omega_3) = y_2$

the mean and standard deviation of the value of the color channels of the pixels in each block. The LUV color space is used here, since moments in this color space are more discriminant than in other spaces, at least for image retrieval [4]. This feature includes 294 components (six for each block).

Color moments are less useful when the blocks contain heterogeneous color regions. Therefore, a global color histogram has been selected as a second color feature. The RGB color space has been subdivided in 64 bins by a uniform quantization of each component in four ranges.

Statistics about the direction of edges may greatly help in discriminating between images depicting natural and man made subjects [18]. To describe the most salient edges we used a 8 bin edge direction histogram: the gradient of the luminance image is computed using Gaussian derivative filters tuned to retain only the major edges. Only the points for which the magnitude of the gradient exceeds a set threshold contribute to the histogram. The image is subdivided into 5×5 blocks, and a histogram for each block is computed (for a total of 200 components).

For their simplicity and satisfactory performance, bag-of-features representations have become widely used for image classification and retrieval [5, 21, 24]. The basic idea is to select a collection of representative patches of the image, compute a visual descriptor for each patch, and use the resulting distribution of descriptors to characterize the whole image. In our work, the patches are the areas surrounding distinctive key-points and are described using the Scale Invariant Feature Transform (SIFT) which is invariant to image scale and rotation, and has been shown to be robust across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination [11]. More in detail, we adopted the implementation described in [20] for both key-points detection and description. The SIFT descriptors extracted from an image are then quantized into "visual words", which are defined by clustering a large number of descriptors extracted from a set of training images [13]. The final feature vector is the normalized histogram of the occurrences of the visual words in the image.

2.3 Combining users

Of course, there is no guarantee that the classes chosen by two different users have a sufficient correlation to make our approach useful. This is why we need several wizards and a method for the selection of those who may help the apprentice organize his pictures. The same argument may be applied to the features as well: only some of them will capture the correlation between the users. Consequently, we treated the features separately instead of merging them into a single feature vector: given a set of pictures labeled by the apprentice, each wizard defines four different classifiers h, one for each feature considered. These classifiers need to be combined into a single classification function that will be then applied to the pictures that the apprentice has not yet labeled.

To combine the classifiers defined by the wizards we apply the multiclass variation of the Adaboost algorithm proposed by Zhu et al. [25]. In particular, we used the variation called Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME). Briefly, given a set $\{(x_i, y_1), \ldots, (x_n, y_n)\}$ of image/label pairs, the algorithm selects the best classifier and assigns to it a coefficient. Different weights are assigned to correctly and incorrectly classified training pairs, and another

classifier is selected taking into account the new weights. More iterations are run in the same way, each time increasing the weight of misclassified samples and decreasing that of correctly classified samples. The coefficients associated to the classifiers depend on the sum of the weights of misclassified samples.

For each iteration the classifier is chosen by a weak learner. The weak learner we defined takes into account all the wizards and all the features. For each wizard u and each of the four features f, a KNN classifier $g_{u,f}$ has been previously trained. Given the weighted training sample, the corresponding mapping functions $\pi_{u,f}$ are computed according to (1); this defines the candidate classifiers $h_{u,f} = \pi_{u,f} \circ g_{u,f}$. The performance of each candidate is evaluated on the weighted training set and the best one is selected. The boosting procedure terminates after a set number T of iterations.

Given an image to be labeled, a score is computed for each class:

$$s_y(x) = \sum_{t=1}^{T} \alpha^{(t)} \chi_{\{y\}}(\overline{h}^{(t)}(x)), \quad y \in Y, \tag{3}$$

where $\overline{h}^{(t)}$ is the classifier selected at iteration t, and $\alpha^{(t)}$ is the corresponding weight. A general schema of classifier creation is shown in Fig. 3. The combined classifier H is finally defined as the function which selects the class corresponding to the highest score:

$$H(x) = \arg\max_{y \in Y} s_y(x). \tag{4}$$

The combined classifier can be then applied to unlabeled pictures. According to [25], the a posteriori probabilities $P(y|x)$ may be estimated as:

$$P(y|x) = \frac{\exp \frac{s_y(x)}{k-1}}{\sum_{y' \in Y} \exp \frac{s_{y'}(x)}{k-1}}. \tag{5}$$

We used the difference between the two highest estimated probabilities as a measure of the confidence of the combined classifier. Unlabeled pictures can then be presented to the user sorted by decreasing confidence.

It should be noted that the output of the classifiers $g_{u,f}$ can be precomputed for all the images of the apprentice. The complexity of the whole training procedure is

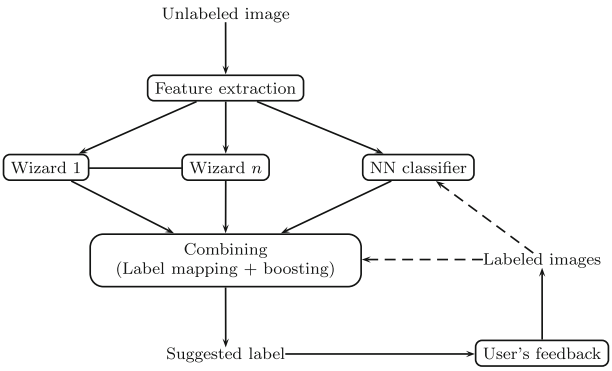

Fig. 3 Simplified schema of the proposed method: for each unlabeled picture, features are computed and passed to the weak classifiers (nearest neighbors and wizards); the output of the weak classifiers is combined into a single suggestion for the input image; the user can confirm the suggestion or make a correction so that the new labeled image can be used to refine the classifiers

$O(nUFT)$, that is, it is linear in the number of labeled pictures n, features considered F, wizards U, and boosting iterations T. The application of the combined classifier to unlabeled pictures may be worked out in $O((N-n)T)$, where N is the number of apprentice's images. Finally, sorting requires $O((N-n)\log(N-n))$. Using the settings described in Section 3, the whole procedure is fast enough, on a modern personal computer, for real time execution and can be repeated whenever a new picture is labeled without degrading user experience.

2.4 Baseline classifiers

In addition to exploiting the information provided by the wizards, we also considered a set of classifiers based on the contents of the apprentice's pictures. They are four KNN classifiers, one for each feature. They are trained on the pictures already labeled and applied to the unlabeled ones. These additional classifiers are included in the boosting procedure: at each iteration they are considered for selection together with the classifiers derived from the wizards. In the same way, it would be possible to include additional classifiers to exploit complementary information, such as camera metadata, which has been proven to be effective in other image classification tasks [1].

The four KNN classifiers are also used as baseline classifiers to evaluate how much our method improves the accuracy in predicting classes with respect to a more traditional approach.

3 Experimental results

To test our method we downloaded from flickr® the images of 20 users. Each user was chosen as follows: (i) a "random" keyword is chosen and passed to the flickr® search engine; (ii) among the authors of the pictures in the result of the search, the first one who organized his pictures into 3–10 sets is selected. In order to avoid excessive variability in the size of users' albums, sets containing less than 10 pictures are ignored and sets containing more than 100 pictures are sub-sampled in such a way that only 100 random images are downloaded. Duplicates have been removed from the albums. The final size of users' albums ranges from 102 to 371, for a total of 3933 pictures.

Unfortunately, some of the selected users did not organized the pictures by content: there were albums organized by time periods, by aesthetic judgments, and so on. Since, our system is not designed to take into account this kind of categorizations, we decided to reorganize the albums by content. To do so, we assigned each album to a different volunteer, and we asked him to label the pictures by content. The volunteers received simple directions: each class must contain at least 15 pictures and its definition must be based on visual information only. The volunteers were allowed to ignore pictures to which they were not able to assign a class (which usually happened when the obvious class would have contained less than 15 images). The ignored pictures were removed from the album for the rest of the experimentation. Table 1 reports the classes defined by the volunteers for the 20 albums considered. Even if our system completely ignores the names used to denote classes, it is interesting to analyze them to understand how the volunteers

Table 1 Summary of the annotation performed by the 20 volunteers

Album	Size	Classes	N. classes
1	328	Animals, artefacts, outdoor, vegetables	4
2	261	Boat, city, nature, people	4
3	182	Close-ups&details, landscapes, railways, portraits&people, sunsets	5
4	251	Buildings, flora&fauna, musicians, people, things	5
5	177	Animals, aquatic-landscape, objects, people	4
6	188	Animals, buildings, details, landscape, people	5
7	151	Arts, city, hdr	3
8	182	Buildings, hockey, macro	3
9	140	Bodies, environments, faces	3
10	227	Animals, beach, food, objects, people	5
11	371	Animals, sea, sunset, vegetation	4
12	168	Animals, flowers, horse racing, rugby	4
13	170	Animals, concert, conference, race	4
14	209	Aquatic, artistic, landscapes, close-ups	4
15	146	Beach, calendar, night, underwater	4
16	134	Animals, family, landscapes	3
17	158	Animals, cold-landscapes, nature-closeups, people, warm-landscapes	5
18	156	Buildings, landscape, nature	3
19	102	Leaves&flowers, men-made, panorama, pets, trees	5
20	234	Microcosm, panorama, tourism	3

For each album are reported the number of pictures and the names given to the classes into which the images have been divided

organized the albums. Considering singular and plural terms as equal, 54 different labels have been used. Volunteers defined a minimum of three and a maximum of five classes. The most frequent labels are “animals” (nine occurrences), “people” (six), “landscape” (five), and “buildings” (four). The number of labels defined by a single volunteer is 44. It is likely that different labels have been used to denote closely related concepts (e.g. “people”, “faces”, “bodies”, “family”). However, it is also possible that different volunteers used the same label to denote different concepts. There is a great variability into the criteria used to annotate the albums; the concepts denoted by the labels range from concrete (e.g. “trees”, “boat”) to very abstract (e.g. “artistic”, “nature”, “things”). However, it should be pointed out that the meaning of labels must be interpreted in the context of the album. For instance, it seems that the label “night” used to annotate album 15 refers to a particular event (a party) and not to the less specific concept of “photos taken at night”. Thirteen volunteers decided to include the “ignored” class, and a total of 193 pictures have been ignored (4.3% of the whole dataset).

To quantitatively evaluate the performance of the proposed method we implemented a simulation of user interaction [6]. This approach effectively allows to evaluate objectively the methodology without taking into account the design and usability of the user interface. The simulation corresponds to the following process:

1. at the beginning all pictures are unlabeled;
2. a random picture is selected and annotated with the correct class;
3. until the whole album is annotated:
   (a) the system is trained on already labeled pictures;

(b) unlabeled pictures are classified;
(c) the picture with the highest classification confidence is selected and annotated with the correct class (i.e. the class assigned to that picture by the volunteer).

As a measure of performance, we considered the fraction of cases in which the class proposed by the system for the picture selected in step 3c agrees with the annotation performed by the volunteer.

The simulation has been executed for the 20 albums considered. Each time an album corresponds to the apprentice and the other 19 correspond to the wizards. Since the final outcome may be heavily influenced by the random choice of the first picture, we repeated the simulation 100 times for each album.

Three variants of the method have been evaluated: (i) using only the KNN classifiers as candidates; (ii) using only wizard-based classifiers; and (iii) using both KNN and wizards. The parameters of the method have been tuned on the basis of the outcome of preliminary tests conducted on ten additional albums annotated by the authors. The number of neighbors considered by the wizards and by the KNN classifiers has been set to 21 and 5, respectively; the number of boosting iterations has been set to 50.

Table 2 shows the average percentage of classification errors obtained on the 20 albums by the three variants of the method. Regardless the variant considered, there is a high variability in performance on the 20 albums, ranging from about 4–60% of misclassifications. Albums 8, 13 and 15 have been organized into classes which

Table 2 Percentage of errors obtained by simulating user interaction on the 20 albums considered

Album	Error rate (%)		
	KNN only	Wizards only	KNN + wizards
1	30.4 (1.5)	28.8 (0.9)	**27.9** (0.9)
2	30.3 (1.3)	33.4 (1.2)	**26.6** (1.8)
3	51.3 (2.1)	47.0 (1.9)	**45.1** (2.1)
4	55.5 (2.0)	55.9 (1.4)	**54.0** (1.8)
5	54.6 (2.4)	54.5 (2.3)	**54.2** (2.2)
6	48.0 (1.9)	48.2 (2.1)	**46.5** (1.9)
7	**24.7** (1.0)	32.8 (1.6)	27.1 (1.9)
8	**12.3** (1.4)	13.2 (1.0)	13.5 (1.2)
9	**43.5** (1.9)	45.4 (2.1)	45.4 (2.1)
10	**31.4** (1.4)	35.9 (1.7)	32.1 (1.5)
11	27.1 (1.1)	27.9 (1.2)	**24.4** (1.3)
12	**20.7** (1.3)	35.7 (1.9)	23.9 (1.7)
13	17.6 (1.2)	18.9 (1.4)	**16.2** (1.0)
14	52.2 (1.9)	51.3 (1.6)	**51.2** (1.7)
15	4.6 (1.4)	10.5 (1.1)	**4.5** (0.7)
16	32.6 (2.1)	30.5 (2.1)	**27.3** (2.1)
17	35.2 (2.3)	39.4 (1.7)	**34.2** (2.0)
18	36.2 (2.1)	34.0 (1.6)	**32.9** (2.1)
19	**57.0** (3.3)	62.5 (3.4)	60.0 (3.6)
20	21.6 (1.4)	21.9 (0.9)	**18.8** (1.2)

The results are averaged over 100 simulations. For each album, the best performance is reported in bold. Standard deviations are reported in brackets

are easy to discriminate and obtained the lowest classification errors. It is interesting to note that these three albums have been the easiest to annotate manually as well (according to informal volunteers' feedback). In particular, albums 13 and 15 have been annotated by the volunteers into classes that are very similar to those defined by the original flickr® users: in both cases the only difference is that two sets have been merged by the volunteers into a single class. The opposite happens for the albums to which correspond the highest classification errors: album 4 originally contained 12 classes, while albums 5 and 19 were organized in 8 classes.

In no case the best result has been obtained using only the wizards-based classifiers. For six albums (1, 3, 5, 14, 16, 18) the wizard-only variant of the method obtained lower errors than the KNN-only variant. It seems that, in the majority of the cases, direct information about image similarity cannot be ignored without a performance loss. The combination of wizards and KNN classifiers outperformed the two other strategies on 14/20 albums. In some cases the improvement is barely noticeable, but in other cases it is significant, with a peak of more than 6% of decrease of misclassifications for album 3. For the other six albums the KNN baseline classifier is the best approach, with a slight improvement over the variant KNN+wizards (a maximum of 3.2% for album 12).

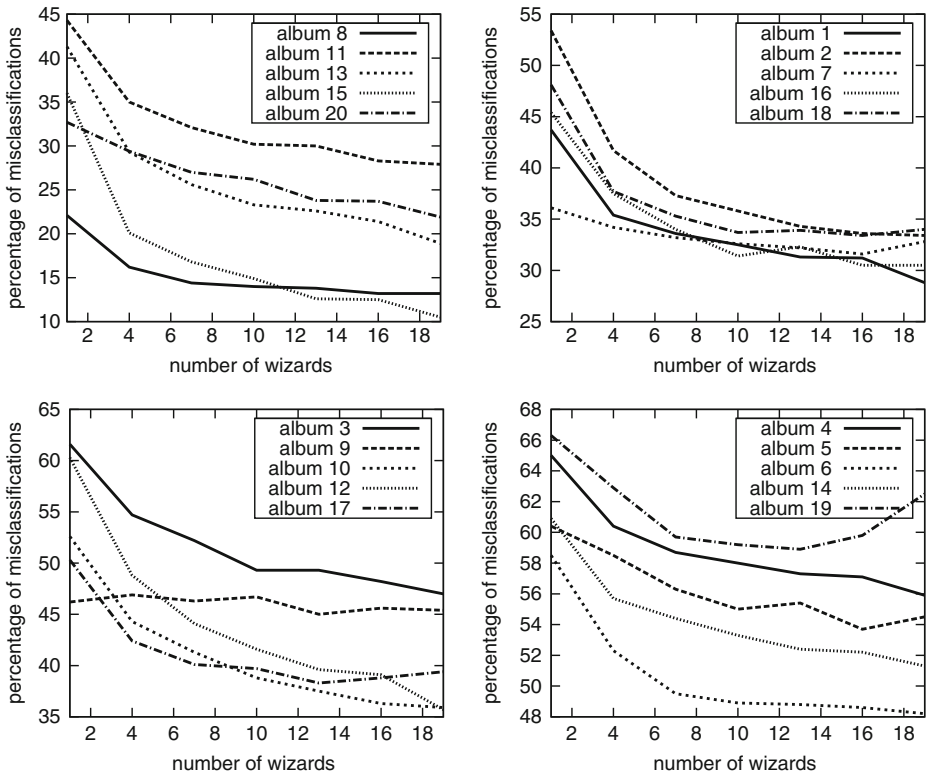

Fig. 4 Percentage of misclassifications obtained on the 20 albums, varying the number of wizards considered. To improve the readability of the plots the albums have been grouped by similar performance

To verify the influence of the number of the wizards on classification accuracy, we repeated the simulations of the wizards-only variant of the method, sampling each time a different pool of wizards. For each album, simulations are performed sampling 1, 4, 7, 10 ,13, 16, and 19 wizards, and each simulation has been repeated 50 times (a different pool of wizard is randomly sampled each time). The plots in Fig. 4 report the results obtained in terms of average percentage of misclassification errors. As expected, for almost all albums, the error rate decreases as the number of wizards increases. The plots suggest that in most cases better performance may be obtained by considering more wizards, in particular for the albums where the lowest errors have been obtained (see the first plot of the figure).

4 Conclusions

In this paper, we described a content-based method the for semi-automatic organization of personal photo collections. The method exploits the correlations, in terms of visual content, between the pictures of different users considering, in particular, how they organized their own pictures. Combining this approach with a KNN classifier we obtained better results (measured on the pictures of 20 flickr® users) with respect to a traditional classification by similarity approach.

We believe that the performace of the method could be improved in several ways. For instance, the method could benefit from the adoption of more powerful machine learning techniques, such as Support Vector Machines. Since the training of wizard-based classifiers is performed off-line, this modification would not prevent real-time interaction. According to experimental results, performance could also be improved by considering a larger pool of wizards.

In this work, we considered the apprentice and the wizards as clearly different characters. We plan to extend our approach to actual photo-sharing communities, where each user would be apprentice and wizard at the same time. However, in order to scale up to millions of wizards (the size of the user base of major photo-sharing websites) a method should be designed for filtering only the wizards that are likely to provide good advices. Moreover, we are considering to exploit additional sources of information such as keywords, annotations, and camera metadata.

An interesting extension of this work would consider group evaluation, such as the storytelling approach of [14]. In this case, the categories of the wizard would be considered as stories, whose characteristics could be measured with suitable features, as done in that paper. Images would then be suggested in such a way that the narrative of the apprentice category would follow that of the wizard's.

Finally, we are investigating similar approaches, based on the correlation between users, for other image-related tasks such as browsing and retrieval.

References

1. Boutell M, Luo J (2004) Bayesian fusion of camera metadata cues in semantic scene classification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 623–630

2. Chen L, Hu B (2003) Face annotation for family photo album management. Int J Image Graphics 3:1–14
3. Das M, Loui A (2003) Automatic face-based image grouping for albuming. In: Proceedings of the IEEE international conference on systems, man and cybernetics, vol 4, pp 3726–3731
4. Furht B (1998) Content-based image indexing and retrieval. In: Handbook on multimedia computing. CRC Press, Boca Raton
5. Grauman K, Darrell T (2005) The pyramid match kernel: discriminative classification with sets of image features. In: Proceedings of the tenth IEEE international conference on computer vision, vol 2, pp 1458–1465
6. Ivory MY, Hearst MA (2001) The state of the art in automating usability evaluation of user interfaces. ACM Comput Surv 33(4):470–516
7. Jaimes A, Benitez A, Chang S-F, Loui A (2000) Discovering recurrent visual semantics in consumer photographs. In: Proceedings of the international conference on image processing, vol 3, pp 528–531
8. Li J, Lim J, Tian Q (2003) Automatic summarization for personal digital photos. In: Proceedings of the fourth international conference on information, communications and signal processing, vol 3, pp 1536–1540
9. Li X, Snoek C, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. In: Proceeding of the first ACM international conference on multimedia information retrieval, pp 180–187
10. Loui A, Savakis A (2003) Automated event clustering and quality screening of consumer pictures for digital albuming. IEEE Trans Multimedia 5(3):390–402
11. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110
12. Mulhem P, Lim J (2003) Home photo retrieval: time matters. In: Proceedings of the international conference on image and video retrieval, pp 308–317
13. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 2161–2168
14. Obrador P, de Olivera R, Oliver N (2010) Supporting personal photo storytelling for social albums. In: Proceedings of ACM multimedia 2010. ACM, New York, pp 561–570
15. Platt J (2000) Autoalbum: clustering digital photographs using probabilistic model merging. In: Proceedings of the IEEE workshop on content-based access of image and video libraries, pp 96–100
16. Shevade B, Sundaram H (2003) Vidya: an experiential annotation system. In: Proceedings of the ACM SIGMM workshop on experiential telepresence, pp 91–98
17. Sun Y, Zhang H, Zhang L, Li M (2002) Myphotos: a system for home photo management and processing. In: Proceedings of the tenth ACM international conference on multimedia, pp 81–82
18. Vailaya A, Jain A, Zhang HJ (1998) On image classification: city images vs. landscapes. Pattern Recogn 31(12):1921–1935
19. Vailaya A, Figueiredo M, Jain A, Zhang H-J (2001) Image classification for content-based indexing. IEEE Trans Image Process 10(1):117–130
20. Vedaldi A (2005) Sift++ a lightweight c++ implementation of sift. http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html. Accessed 01 Oct 2011
21. Wallraven C, Caputo B, Graf A (2003) Recognition with local features: the kernel recipe. In: Proceedings of the ninth IEEE international conference on computer vision, vol 1, pp 257–264
22. Wenyin L, Sun Y, Zhang H (2000) Mialbum—a system for home photo managemet using the semi-automatic image annotation approach. In: Proceedings of the eighth ACM international conference on multimedia, pp 479–480
23. Zhang L, Chen L, Li M, Zhang H (2003) Automated annotation of human faces in family albums. In: Proceedings of the eleventh ACM international conference on multimedia, pp 355–358
24. Zhang J, Marszalek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vision 73(2):213–238
25. Zhu J, Rosset S, Zou H, Hastie T (2005) Multiclass adaboost. Technical report, Stanford University. Available at http://www-stat.stanford.edu/~hastie/Papers/samme.pdf. Accessed 01 Oct 2011

Claudio Cusano is a post-doc researcher at DISCo (Department of Information Science, Systems Theory, and Communication) of the University of Milano-Bicocca, where he took his PhD in Computer Science. Since April 2001 he has been a fellow of the the ITC Institute of the Italian National Research Council. The main topics of his current research concern 2D and 3D imaging, with a particular focus on image analysis and classification, and on face recognition.

Simone Santini (M '98) received the Laurea degree from the University of Florence, Italy, in 1990, the MSc and the PhD degrees from the University of California, San Diego (UCSD) in 1996 and 1998, respectively. In 1990, he was a visiting scientist at the Artificial Intelligence Laboratory at the University of Michigan, Ann Arbor, and, in 1993, he was a visiting scientist at the IBM Almaden Research Center. He has been a project scientist in the Department of Electrical and Computer Engineering, UCSD, and a researcher at Praja, Inc. Since 2004 he is an associate professor at the Universidad Autónoma de Madrid (Spain). His current research interests are interactive image and video databases, behavior identification and event detection in multisensor stream, query languages for event-based multimedia databases, and evaluation of interactive database systems.