Semantic Multimedia: 4th International Conference on Semantic and Digital Media Technologies, SAMT 2009 Graz, Austria, December 2-4, 2009 Proceedings. Lecture Notes in Computer Science, Volumen 5887. Springer, 2009. 28-39.

**DOI:** http://dx.doi.org/10.1007/978-3-642-10543-2_5

# On the coöperative creation of multimedia meaning

Claudio Cusano and Simone Santini and Raimondo Schettini

Università di Milano Bicocca, Milano, Italy
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Università di Milano Bicocca, Milano, Italy

**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 make a correction. 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. A quantitative evaluation of the proposed approach is proposed based on pictures of a set of members of the flickr® photo-sharing community.

## 1 Introduction

The process of signification, that semantic computing tries to unravel using formal means, already extremely complex in the case of text, acquires new dimensions and nuances in the case of multimedia data. An image, per se, doesn't have any meaning, being just a recording of a certain situation that happend to unfold in front of a camera at a certain point in the past. Its only inherent meaning can be described as the Barthesian *ca-a-été*: the thing that is represented happened in the past. But, of course, many things happened in the past that were not recorded in images, and the meaning of an image is related to a decision: the decision to record certain things and not others. Photos are not taken higgledy-piggledy, but according to certain discoursive practices that depend on the purpose of the picture and on the community in which they are taken.

Taking a picture in order to convey a meaning is an activity that follows certain socially-dictated rules. These rules are with us from the beginning of our picture-taking life, and we follow them more or less unconsciously. When we are on vacation, we take mostly pictures of stereotypically happy moments, often in front of the same sceneries and monuments, and we avoid certain themes (sexual situations, for example). Often, these practices tell us about the meaning of a picture more than the contents of the picture itself.

These observations, schematic and superficial as they may be, point to the impossibility of creating a *semantic* image classification system based only on the contents of the images. Semantic classification entails the division of the image space along semantic lines, and these lines depend crucially on the discoursive practices that preside

image acquisition and on the interpretative practices of the community to which the images are directed.

In this day and age, fortunately, a lot of information about community practices is available in a conformation that affords formalization. Thanks to the emergence of on-line communities, community practices can be understood by analyzing the way people organize their data on the internet. Our current work aims at using this structural information for understanding the semantics of images and, in a broader view, for understanding the process of signification in multimedia.

The system that we present in this paper is a simple outcome of this activity, and it aims at helping people in a task that, with the advent of digital cameras, has become fairly common: to classify personal photographic pictures, dividing them into thematically organized folders. The criteria that preside this organization are, of course, 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. However, the discoursive practices that preside to this classification are, to a certain extent, common to all users. That is, all said and done, people are not that original. Nor could they be: in the internet communit era, photos, and their classification scheme are communication means, and communication can only work through a shared code. Classification is part of a semiotic system, and must have some degree of uniformity and predictability, at least within the community in which the communication is done.

Faithful to the principles of community based semantic creation, we try to use the collective wisdom of the community in order to suggest to one of its member possible ways of classification. Briefly, when a person (we call this person the *apprentice*) start putting photos into carpets, the system will look at other users of the community and at the classifications they made. Members who agree with the classification made by the apprentice (yclept the *wizards*) will be used as classifiers to propose a classification of the apprentice's unclassified images.

We can see a system like this under two possible lights. On the 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 upfront, 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 users. 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.

Commercial systems for the management and classification of personal photos rely essentially on manual annotation, and their only distinguishing trait is the interface that they use to make annotation as rapid and convenient as possible. Research prototypes take a more ambitious view, and try to provide tools for automatic or (more often) semi-automatic classification. A prototype system for home photo management and processing was implemented by Sun et al. [13]. 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). The system developed by Wenyin et al. [17] 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, and the keywords with higher frequencies in these images are used to annotate the new images. Mulhem and Lim proposed the use of temporal events for organizing and representing home photos using structured document formalism [9]. Shevade and Sundaram presented an annotation paradigm that attempts to propagate semantic by using WordNet and low-level features extracted from the images [12]. 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 to the automatic organization of photo albums consists in the application of clustering techniques, grouping images into visually similar sets. Some manual post-processing is usually required to modify the clusters in order to match user's intended categories. Time information 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 [11]. Li et al. exploited time stamps and image content to partition related images in photo albums [6]. 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. [5]. 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.

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 [2].

The idea of exploiting user correlation in photo sharing communities has been investigated by Li et al. [7]. 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 The system

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 process the images is not restricted. A suitable user interface will allow a rapid label confirmation, and a quick and easy organization of the photo album.

### 2.1 Correlation within the community

One of the difficulties of assisted album organization is that, at the beginning, we lack information on the criteria that the user is going to apply in 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 commom characteristic, at least at the semantic level. 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 (the *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. Simply (maybe overly so) put: 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 pictures are used as a training set in order to train a classifier that implements a classification function $g : X \rightarrow \Omega$ from the feature space $X$ into the set of wizard 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 categories different from those of the wizard. 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 integratation 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 $\omega_i$; then $\pi$ 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, $\pi$ maps a class $\omega$ of the wizard into the class of the apprentice that maximizes the cumulative weight of the images that $g$ maps back into $\omega$. If no apprentice image belong $\omega$ we define $\pi(\omega)$ to be the class of maximal total weight.

If we interpret $w$ has a misclassification cost, our definition of $\pi$ 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 $\chi$ denotes the indicator function ($\chi_A(x) = 1$ if $x \in A$, 0 otherwise). 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, $h$ shows a useful property: the part defined by $g$ is independent of the apprentice, so that it can be computed off-line, allowing the adoption of complex machine learning algorithms such as SVMs, neural networks, and the like; the part defined by $\pi$, instead, can be computed very quickly since it 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 which 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. We use 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. [14], we divided each image into $7 \times 7$ blocks and computed the mean and standard deviation of the value of the color channels of the pixels in each block. This feature is made of 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 [15]. 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 $5 \times 5$ blocks, and a histogram for each block is computed (for a total of 200 components).

Bag-of-features representations have become widely used for image classification and retrieval [18, 16, 3]. The basic idea is to select a collections 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 robust vis-a-vis a substantial range of distortions [8]. 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 [10]. 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 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 will then 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. [19]. 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 assign 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 depends on the sum of the weights of misclassified samples.

For each iteration the classifier is chosen by a weak learner that 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. 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 [19], 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 $O(nUFT)$, 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 the user's 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 to 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.

To quantitatively evaluate the performance of the proposed method we implemented a simulation of user interaction [4]. This approach effectively allows to evaluate objectively the methodology without taking into account the design and usability of the user interface. 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; iii) using both KNN and wizards.

**Table 1.** Summary of the annotation performed by the 20 volunteers. For each album are reported the number of pictures and the names given to the classes into which the images have been divided.

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

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% to 60% of misclassifications. Albums 8, 13, and 15 have been organized into classes which 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

**Table 2.** Percentage of errors obtained by simulating user interaction on the 20 albums considered. The results are averaged over 100 simulations. For each album, the best performance is reported in bold. Standard deviations are reported in brackets.

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

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).

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 Figure 1 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.

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.

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

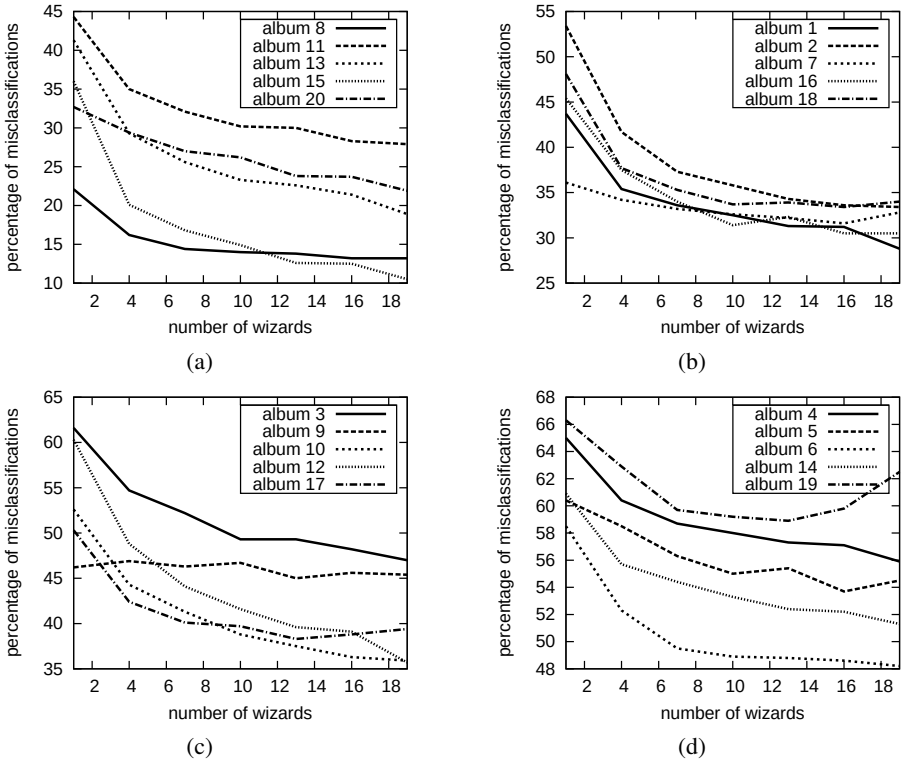


**Fig. 1.** Percentage of misclassifications obtained on the 20 albums, varying the number of wizards considered. To improve the readability of the plots the albums have been grouped by similar performance.

## References


1. M. Boutell and J. Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 623–630, 2004.
2. M. Das and A. Loui. Automatic face-based image grouping for albuming. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3726–3731, 2003.
3. K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 2, pages 1458–1465, 2005.

4. M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516, 2001.
5. A. Jaimes, A. Benitez, S.-F. Chang, and A. Loui. Discovering recurrent visual semantics in consumer photographs. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 528–531, 2000.
6. J. Li, J. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *Proceedings of the Fourth International Conference on Information, Communications and Signal Processing*, volume 3, pages 1536–1540, 2003.
7. X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proceeding of The first ACM International Conference on Multimedia Information Retrieval*, pages 180–187, 2008.
8. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
9. P. Mulhem and J. Lim. Home photo retrieval: Time matters. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 308–317, 2003.
10. D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168, 2006.
11. J. Platt. Autoalbum: clustering digital photographs using probabilistic model merging. In *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 96–100, 2000.
12. B. Shevade and H. Sundaram. Vidya: an experiential annotation system. In *Proceedings of the ACM SIGMM Workshop on Experiential Telepresence*, pages 91–98, 2003.
13. Y. Sun, H. Zhang, L. Zhang, and M. Li. Myphotos: a system for home photo management and processing. In *Proceedings of the Tenth ACM International Conference on Multimedia*, pages 81–82, 2002.
14. A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, Jan 2001.
15. A. Vailaya, A. Jain, and H. J. Zhang. On image classification: city images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935, 1998.
16. C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 1, pages 257–264, 2003.
17. L. Wenyin, Y. Sun, and H. Zhang. Mialbum — a system for home photo managemet using the semi-automatic image annotation approach. In *Proceedings of the Eighth ACM International Conference on Multimedia*, pages 479–480, 2000.
18. J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.
19. J. Zhu, S. Rosset, H. Zou, and T. Hastie. Multiclass adaboost. Technical report, Stanford University, 2005. Available at http://www-stat.stanford.edu/ hastie/Papers/samme.pdf.