Challenges in Content-Based Image Indexing of Cultural Heritage Collections

David Picard, Philippe-Henri Gosselin, Marie-Claude Gaspard

Challenges in Content-Based Image Indexing of Cultural Heritage Collections

David Picard[(1)], Philippe-Henri Gosselin[(1)] and Marie-Claude Gaspard[(2)]

***Abstract*—Cultural heritage collections are being digitized and made available through online tools. Due to the large volume of documents being digitized, not enough manpower is available to provide useful annotations. Is this paper, we discuss the use of automatic tools to both automatically index the documents (*i.e.*, provide labels) and search through the collections. We detail the challenges specific to these collections as well as research directions that have to be followed to answer the questions raised by these new data.**

I. Introduction

DIGITIZATION of cultural heritage collections has recently become a topic of major interest. Large campaigns of digitization are launched by several institutional and private entities to allow instant access to billions of documents. Thanks to these new portals, anyone can see these collections that were usually stored in archives under restricted access[1]. These campaigns are both rich in the number of digitized documents and in the number of targeted subjects. Together, these points make them appealing as a research topic and as a new tool for the final user.

While the idea of an open access to digital copies of all kinds of historical contents is very appealing, the size and variety of these new data lead to a wide range of new problems. In particular, the pace at which historical artifacts are digitized greatly exceeds the manpower needed to manually index them. By indexing, we mean labeling using carefully chosen keywords for all documents so as to ease the search through the entire collection. The online and open aspect of digitized collections raises new questions with respect to the uses emerging from the size of available data as well as from the variety of users searching them.

In this paper, we consider automatic labeling and interactive search challenges. In automatic labeling, the goal is to automatically infer a set keywords for each newly digitized artifact. The thesaurus of all possible keywords can be very large and can contain concepts with varying semantic degrees. The main goal of automatic labeling is to ease the work of specialists searching throughout the entire collection by querying very precise keywords. In interactive search, results of a query (either starting from an example or from a keyword) are graphically shown to the user. These results can be refined thanks to user feedback, for instance through the highlighting or the removal of some elements. The goal of interactive search is to find documents that can not be retrieved using keywords in a minimum amount of interaction.

The remainder of this paper is organized as follows: In the next section, we present the specificity of cultural heritage collections. We highlight their differences with other image collections currently used in image processing and computer vision communities. To illustrate our points, we present the Bilbiothèque nationale de France (BnF) image collection. Then, in section III, we present a brief survey of current techniques for content-based image indexing. In section IV, we propose a baseline evaluation on the BnF collection for both automatic labeling and interactive search. Finally, we conclude and discuss the open questions.

II. Cultural heritage digital image collections

This section examines to problem of indexing cultural heritage collections. For this purpose, we first present the example of a labeled subset of the Bibliothèque Nationale de France image collection. Based on this presentation, we then detail the expected difficulties that are to be tackled when indexing such collections.

A. The BnF image collection

Examples of images digitized by the Bibliothèque Nationale de France and their associated labels are shown in Figure 1. The currently online collection contains around 275 000 images. Around 14% of them are labeled with one or several keywords. The images are pictures from any kind of cultural heritage artifact such as paintings, coins, tapestry, manuscript, *etc*. The corresponding labels vary from very generic terms (*e.g.*, *Animal representation* for the top left image), to very specific ones (*e.g.*, *Ptolémée V* for the bottom left image).

The images and their keywords can be accessed on the online website of the BnF[2]. The advance search feature can be used to reveal the hierarchical architecture of the thesaurus, as well as the statistics of occurrences of the keywords. The main level of the hierarchy is divided in broad categories stating the image acquisition geometry, the picture genre, its time and geographical information and its subject in terms of objects and people.

In this paper, we focus on a small subset of about 4000 labeled images. In Figure 2, we show the distribution of the

D. Picard and P.-H. Gosselin are with ETIS / ENSEA - Univ. Cergy-Pontoise - CNRS, F-95000 Cergy, {picard,gosselin}@ensea.fr

M.-C. Gaspard and S. Petratis are with Bibliotheque nationale de France - Departement de la reproduction, Quai François Mauriac, F-75706 Paris, marie-claude.gaspard@bnf.fr

[1]See for examples the online collection of the Library of Congress (http://www.loc.gov/pictures), or that of the Metropolitan Museum of Art (http://www.metmuseum.org/collection/the-collection-online).

[2]http://images.bnf.fr

bord de mer, chèvre, fortification, représentation animalière, représentation scientifique

cartouche, corde, cuir ornemental, église, moulin à eau, Namur (province), oiseau, ornementation, paysage, perle ornementale, personnage, rivière, route, tour, village

Arsinoé III, reine d'Egypte, couronne, de profil, diadème = bandeau, draperie, en buste, épi, femme, grènetis, homme, lance, portrait, Ptolémée V, roi d'Egypte, sceptre, (0323-0031 av. J.-C.) Epoque hellénistique, Grec

âne, lion, renard, représentation animalière

Fig. 1. Examples of images and their corresponding labels (in French) from the BnF collection. Observe the variety in the keywords, from generic (*Animal representation* in the top left image) to very specific (*Ptolémée V* in the bottom left image) topics.

number of labels per image. As we can see, only around 120 images have only one label. The vast majority of images have between 2 and 15 labels.

In Figure 3, we show the distribution of the number of images per class. There are many classes with only one corresponding image, and these classes usually correspond to very specific labels. Examples of classes with 1 image are: *Philippe IV le Bel*, *époque Louis XIII*, *Papouasie-Nouvelle-Guinée*, *la guerre*, *électricité*, *marteau d'armes*, *générosité*. The greater the number of images per class, the fewer the number of classes (the peak at 100 in Figure 3 refers to the cumulative tail of the distribution). This distribution gives hints to the specificity of the classes found in cultural heritage collections.

Contrary to the RKD challenge presented in [13], indexing the BnF is much more complex. The tasks in the RKD challenge are to predict the author, the type of work, the material and the time of creation. The number of samples available for these classes fairly outnumbers the one for the classes in the BnF. Furthermore, apart from the author identification, the classes of the RKD challenge are based on physical properties and not on semantic visual content. Nonetheless, the RKD provides a web search engine allowing users to search its entire digitized collection with keywords[3]. Althought these keywords are highly semantic like those of the BnF, they were not retained for the RKD challenge. Therefore, there is a need for a public dataset that encompasses the full difficulty of indexing cultural heritage collections and allowing researchers to assess their tools.

[3]http://www.rkd.nl

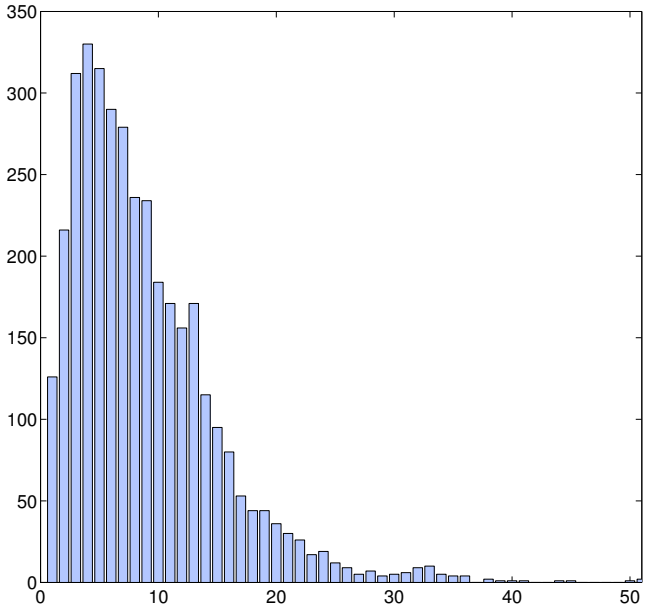

Fig. 2. Histogram of the number of labels per image in the BnF collection. The last bin corresponds to all images with more than 50 labels.

B. Open questions

Compared to the generic image collections used in current computer vision benchmark, labeling cultural heritage collection is much more difficult, due to the wide range of expected labels and to the very specific knowledge required to understand them. As we can see in Figure 1, some of the labels are sufficiently common to be inferred by everyone,

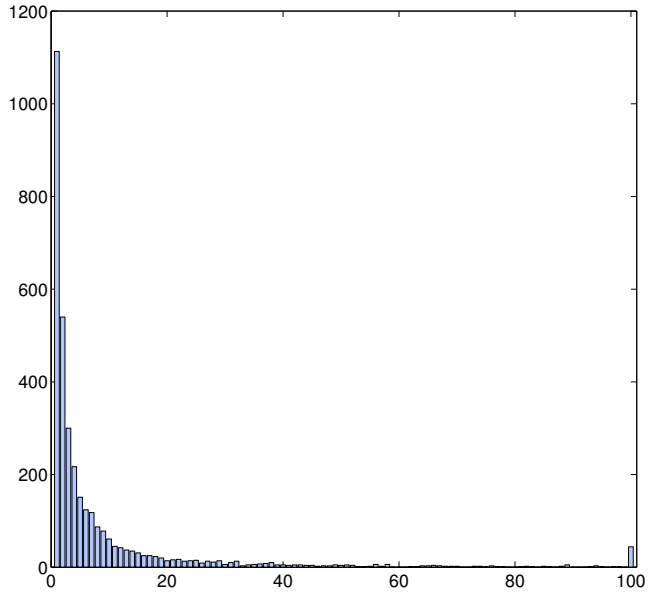

Fig. 3. Histogram of the number of images per class in the BnF collection. The last bin corresponds to all classes with more than 100 images.

but other ones require specific knowledge in history or in material science. To better evaluate the difficulty of labeling such collections, we propose to divide the labels in several domains:

- Visual characteristic (shape, color, etc)
- Semantic content (objects within the image, category of art: portrait, landscape, manuscript, etc)
- Physical properties (canvas, marble, wood, paper, etc)
- Historical information (production period, name of a character, style, etc)
- Geographical information (geographic name, towns, regions, map, etc)

In this list, only the visual characteristic is effectively tackled by current content-based indexing methods. Semantic content has currently promising results thanks to the recent development in computer vision. There is, as far as we know, very few work on the remaining domains.

The main difficulty induced by this list is that methods need to be based on specific properties of the signal to perform well on some very specific classes. For example, predicting the type of paper might use different image characteristics (different scales, different modalities) than predicting the style. Designing a generic system that can automatically select the signal properties adapted to each specific class, without being a complex combination of ad hoc methods, is the main challenge of the proposed tasks.

The second problem induced by the wide class diversity is the scarcity of labeled samples. Indeed, some labels have by nature very few examples (*e.g.*, "King Louis XI of France") and very few of them are labeled. Unlike generic image collections, as the cultural heritage image collections grow, it is unlikely that the number of samples per class increases much for the vast majority of classes. Instead, it is more likely that the number of classes grows, while the number of samples per class remains almost constant. This arises challenging learning problems, since we have to build a large set of similarity measures and classifiers with very few samples.

III. Content-based image similarity

The main technical challenge associated with the indexing of image collections is to be able to compute a numerical representation of each image and its associated similarity measure. This similarity measure aims at being as close as possible to human visual perception.

Generally speaking, the design of such representations and their associated metric is not an easy task as it has to bridge the semantic gap [22]. Several families of similarity measure have been defined in recent years, in many cases with a very specific goal in mind. Indeed, the design of such measures is highly dependent on the application (*e.g.*, object recognition, scene understanding), since it allows the incorporation of prior knowledge which usually boosts the performances. In the following, we list most of the existing families of content-based image similarity, detailing their original application, and showing some of their adaptations to art collections.

A. Global approaches

Historically, the first approaches in defining a similarity between images were using a global index. A global index means that the numerical representation computes statistics on the properties of the signal at the scale of the whole image. Simple examples of such techniques include color histograms [23] or texture histograms [12].

With respect to cultural heritage images, global descriptors such as GIST have been used for image alignment and registration [18]. However, they suffer the same drawbacks as for the general image labeling and retrieval tasks: They are not able to handle classes discriminated by local visual properties. In particular, when the goal is to retrieve a specific instance of an object (*e.g.*, a specific Roman emperor coin from the category coins), it becomes obvious that statistics at the image level are not sufficient to discriminate this specific instance from the other of the same category.

B. Local descriptors

To solve the precision problem of global descriptors, local keypoint matching techniques have successfully been proposed. The key idea behind the keypoint matching techniques is to select a set of salient regions in the image (denoted *"Region of Interest"* or ROI), to compute a description of the content of the region and then to perform a pairwise matching of the keypoint descriptions between two images. The more keypoints match between images, the more they are considered similar.

The ROI detection step is based on salience measures like corner detectors [7] or blob detector [11]. A good overview of the keypoint detection techniques can be found in [14]. Recently, it has been found that a dense extraction of keypoints leads to even better similarity measures, at the cost of a more complex matching step.

The ROI description is in some sense very similar to that of the global indexes, except that it is only computed on a small region of the image. The most used descriptors in current systems are SIFT [11] or HOG [5] which basically compute an histogram of the gradient orientations in cells spanning the ROI. With such descriptors, the shape of the edges in the ROI is encoded.

Once descriptors are extracted, measuring the similarity between two images is akin to counting the number of matching pairs of descriptors. Given a descriptor d of the first image, its nearest neighbor $1\mathrm{NN}(d)$ in the second image is considered as a match if their distance is less than a threshold relative to the distance with the second nearest neighbor $2\mathrm{NN}(d)$:

$$d(x, 1\mathrm{NN}(x)) < \lambda d(x, 2\mathrm{NN}(x)) \tag{1}$$

with typically $\lambda = 0.6$ [11]. The assumption is that a given descriptor in the first image has a unique corresponding descriptor in the second image; together these form the closest pair.

To extend this matching scheme to an entire image collection, we consider the set $B = \cup_i B_i$ of all descriptor sets B_i of all images i in the collection. Then, for each descriptor from the query q, its k nearest neighbors are retrieved from the entire collection. Every image receives as many votes as nearest neighbors it contains. Votes are summed up for all descriptors of the query, and the image with the highest number of votes is the most similar:

$$s(q, i) = \sum_{d \in B_q} (\mathrm{kNN}_B(d) \cap B_i) \tag{2}$$

However, such matching scheme is unable to scale with larger collection and larger sets of extracted descriptors. To run scalable searches, most of accelerating schemes are based on approximate nearest neighbor search in high dimensional spaces, such as Inverted Files [21] or Locality Sensitive Hashing [6].

With respect to cultural heritage collections, the main assumption driving local descriptor matching is relevant for duplicate or near duplicate retrieval. Searching for a seal, a coin or a specific printed pattern are clear examples where the assumption holds. More semantic queries, such as author identification or time estimation can also be tackled using these approaches, depending on the scale of the images. For example, the brush stroke of a painter doing a specific pattern such as the ear of a character is a highly distinctive region of interest that can be matched in another painting. In [4], the authors proposed a matching scheme to perform object detection in paintings while training on natural images, and show matching local parts of an object improve the performances. However, they stay with very generic categories such as *dog* or *chair*.

The main challenge of these approaches in art related collections is the selection of the right detector/descriptor couple to obtain satisfying results for a specific application. Unfortunately, no generic local detector/descriptor couple is able to tackle all the similarities that can be defined in such collections due to the large variability of scale (from canvas threads to scene layout) and materials (parchment, canvas, marble, metal, *etc*).

C. Aggregating methods

While local descriptors matching usually leads to very high performances, accelerating scheme are not sufficiently efficient to deal with very large collections. In particular, since all descriptors are kept, the amount of data to be stored grows with the number of extracted descriptors per image and the number of different modalities.

To overcome this problem, aggregating methods have been proposed to reduce the representation of an image from a set to a single vector. To perform the aggregation, most methods use a dictionary D of M prototypical descriptors $D = \{\mu_c\}_{1 \leq c \leq M}$. The set of descriptors B_i of an image i can then be described in term of statistics over D. The first of such method, called *Bag of Visual Words* (or BoW), assigns each descriptor of the image to its closest entry in the dictionary and computes the histogram of such assignments [21]. Formally, the assignment corresponds to a quantization function q that returns a vector filled with zeros except for a 1 at the component corresponding to the closest prototype:

$$q(x) = [\delta_{ik}]_i,\ k = 1\mathrm{NN}(x) \tag{3}$$

The signature is simply the sum of all these vectors:

$$x_i = \sum_{x \in X} q(x) \tag{4}$$

The interesting part of the BoW method is that it corresponds to a matching function between two sets of descriptors, where a match is found if and only if the two descriptors are assigned to the same prototype:

$$s_{BoW}(q, i) = \left(\sum_{d \in B_q} q(d) \right)^\top \left(\sum_{p \in B_i} q(p) \right) \tag{5}$$

$$= \sum_{d \in B_q} \sum_{p \in B_i} \delta_{kl},\ k = 1\mathrm{NN}(d),\ l = 1\mathrm{NN}(p) \tag{6}$$

More recently, refinement in the encoding of the descriptors have been proposed. For instance, instead of simply quantizing each descriptor to its closest prototype, sparse coding and related methods [25] propose to view the encoding as a constrained reconstruction problem:

$$q(x) = \mathrm{argmin}_\alpha \|x - D\alpha\|^2 + \lambda \Omega(\alpha) \tag{7}$$

where $\Omega(\alpha)$ is a regularizer, typically the ℓ_1 norm to enforce a sparsity pattern in the coefficients α or a locality constraint to ensure descriptors are encoded by nearby prototypes in the case of Locality constraint Linear Coding [25]. The incentives behind such encoding schemes are that the reconstruction term reduces information loss when compared to hard quantization approaches. Furthermore, the aggregation of codes introduces a minimum averaging effect due to the sparsity pattern in the codes. Since D is an overcomplete dictionary and using the generalized Parseval identity, there is a relationship between the dot product of two descriptors in the descriptors space and the dot product of their coding coefficients. As a consequence,

the dot product of two signatures is related to a matching scheme where the descriptors are compared using the dot product.

The idea of using a matching scheme that can be linearized into a single vector has thus been proposed in several methods. In Vectors of Locally Aggregated Descriptors (VLAD) [8], the authors assign each descriptor to its closest prototype, and then to encode the differences between the descriptor and the prototype:

$$q(d) = [\delta_{ik}(d - \mu_i)]_i,\ k = \mathrm{1NN}(d) \tag{8}$$

The aggregation is simply the sum of all codes, like in BoW. The corresponding matching scheme compares only descriptors assigned to the same prototype and computes the match using the dot product of the descriptors centered on their respective prototypes:

$$s_{VLAD}(q, i) = \sum_{d \in B_q} \sum_{p \in B_i} \delta_{kl} \langle d - \mu_k, p - \mu_l \rangle, \tag{9}$$

$$k = \mathrm{1NN}(d),\ l = \mathrm{1NN}(p) \tag{10}$$

By looking at VLAD, we can clearly see the bridge between matching schemes and the comparison of different statistics over D. In that sense, BoW is a piecewise constant matching scheme and corresponds to a zero order statistic, while VLAD is a piecewise linear matching scheme and corresponds to a first order statistic.

To improve the discriminative capability of the similarity measure, higher orders have been proposed. In particular, Fisher Vectors [16] consider second order information by computing the dictionary as a Gaussian mixture model. Then, it assigns the descriptors to all components of the mixture proportionally to their likelihood. Finally, it computes the first and second order moments of the descriptors with respect to each component. Using a hard assignment, Vector of Locally Aggregated Tensors (VLAT) [17] computes higher order moments using the tensor products of descriptors. The final signature is then the concatenation of all orders. The authors show that the dot product between the signatures is equivalent to a matching scheme using the dot product between the descriptor raised to the power t, which in turn is a approximation of a Gaussian matching kernel between the descriptors using a Taylor expansion of order t.

Considering cultural heritage images, aggregating approaches are likely to generalize in the labeling task, as shown with Fisher Vectors in [3]. In interactive search, their relation to keypoint matching is also likely to obtain a good accuracy thanks to the discriminative power of such schemes.

D. Deep architectures

In contrast with the two steps of the local descriptor aggregation approaches, deep architectures stacking several layers of encoding have also been proposed. While deep neural networks have been proposed for a long time [10], their recent success in generic image classification benchmarks revived recent development of such methods.

The greatest advances were made with the use of Convolutional Neural Networks (CNN) which alternate layers of convolutional filters and layers of pooling operations [9]. The weights of the neurons on the convolution layers correspond to the coefficient of filters and can be trained in two steps. First, a pre-training step minimizing the reconstruction error trains a preliminary set of auto-encoding filters. Second, the filters coefficients are tuned by back propagation of the classification error. This second step helps locate the right combination of filters.

This repetition of convolution, non-linear operation and pooling bears some similarity with the multiscale analysis performed by wavelets. Indeed, the authors of [1] propose a deep architecture composed of layers of wavelet filters combined with a non linear operator (modulus) to obtain invariance to certain transforms.

Although most CNN are used directly in classification tasks, it was empirically shown that the layers before the classification provide very good image representation that can be used in almost any image similarity related task [19]. It has also been recently shown that stacking many layers [20] further improves the performances, which raises the question of the structure rules to follow when designing a deep neural network.

Since CNN provide a strong baseline for image features, they were already used with paintings in [3]. The authors propose to classify paintings using a training set of natural images taken from Google images. They achieve surprisingly good performances considering the discrepancy between the objects in natural images and their depiction in paintings. However, the experiments are limited to a small number of generic classes (*e.g.*, *boat*, *car*, *horse*), and can not be easily extended to the very specific classes we consider in this work.

IV. Applications and experiments

In this section, we present baseline results on the BnF collection to show the complexity of the challenging applications. We set up a benchmark with rigorous evaluation procedure allowing to compare different visual features[4]. First, we present results on automatic labeling, and then on interactive search.

A. Automatic labeling

In these experiments, we only considered classes with more than 10 images to be able to compute relevant statistics. We found 569 classes corresponding to this criterion, with the following repartition: Semantic (459), visual (62), historical (26), geographical (14), physical (8). We used a standard classification approach consisting in a single mid-level feature per image combined with a linear SVM. This setup compares to most pipeline used in current academic challenges. We trained the SVM on a 1 vs rest mode for each class. Using 5-fold cross-validation, we computed the Average Precision (AP) for each class.

We compare 3 different types of mid-level features, namely Fisher Vector [16], CPVLAT [15] and deep CNN based features taken from the penultimate layer of CNN-s in [2].

[4]The benchmark can be downloaded at `http://perso-etis.ensea.fr/ picard/bnf_bench/`

	Fisher Vectors	CPVLAT	CNN-s
Semantic	16.9	14.4	16.7
Visual	35.4	25.2	32.1
Historical	18.8	16.8	20.2
Geographical	34.7	28.1	31.3
Physical	31.2	23.4	28.7
Combined	27.4	21.6	25.8

TABLE I
RESULTS (% MAP) OF THE LABELING TASK FOR DIFFERENT FEATURES.

These 3 different features allow to assess the behavior of different families of methods, in particular probabilistic models for Fisher Vector, Keypoint Matching for CPVLAT and Deep Neural Networks for CNN-s. Remark that contrarily to other methods, CNN-s are trained on a much larger dataset (ImageNet) not related to cultural heritage.

We show in Table I the results of each feature with respect to the different categories. Fisher Vectors almost consistently outperforms the other methods which shows that classical computer vision methods provide a strong baseline for visual similarity even with a large variety of classes. As we can see, the combined mAP is less than 28%, which shows the complexity of the task. Semantic is the most difficult category, whereas Visual is the easiest, although these categories are the ones containing most classes and thus most samples. We show in Figure 4 detailed results for some of the best and worst classes for each feature. While it is easy to understand why some classes are difficult, like *drum* or *whip* which are blurry tiny details combined with few training samples (between 10 and 15), it is worth remarking that most of the easy classes obtain good results mainly for statistical reasons. For instance, all the images with the *Danish* label are from the same dataset of sketches representing the Danish army during the 18th century. Likewise, the *sewer* class contains only maps of the sewers of Paris which are visually very similar.

B. Interactive search

In interactive search, we consider the scenario where a user is searching for a subset of the collection corresponding to a specific concept. This can be the case when a new class is inserted in the thesaurus for example. In that case, the user wants to retrieve all the images belonging to the concept with the minimum amount of interaction with the system (*i.e.*, the minimum number of clicks). A typical session is as follows: Starting from a single image belonging to the class, we interactively select five images, label them and add them to the pool of known images, retrain the classifier and present the results for the next round of interaction until 50 images are labeled.

To evaluate performances in this setup, we compared the performances of two strategies of interaction by measuring the mAP against the number of labeled images (averaging 10 sessions per class). The "*BestSample*" strategy selects the most relevant sample as evaluated by the current classifier (*i.e.*, $\max_x f(x)$ with f the current classifier), while the "*Simple*" strategy [24] selects the sample closest to the margin (*i.e.*, $\min_x |f(x)|$). Note that the labeled images are counted when computing the mAP, which biases the results compared to the classical labeling task. However, counting the labeled samples is coherent with the end-user application which should provide all correct results to the user, including those seen during the interaction.

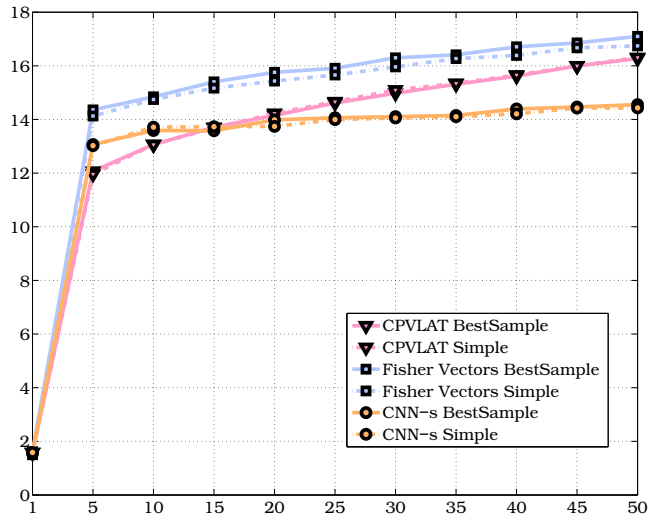

Fig. 5. mAP against the number of labeled images for 2 active strategies.

We show in figure 5 the mAP against the number of labeled images. As we can see, both strategies perform about the same regardless of the features, with a slight advantage for the BestSample strategy. This can easily be explained by the nature of our data, where most classes are very small and diverse. In such case, samples close to the margin are likely to be negative ones, mainly due to the low appearance probability of the relevant class. Fisher Vectors also outperform other features in this task. Remark that contrarily to the labeling task, CPVLAT offers better performances than CNN-s, which means this feature is more able to discriminate and less to generalize, which is consistent with the retrieval task. This is probably due to the strong relation of CPVLAT with keypoint matching schemes.

Furthermore, it is interesting to note that the best mAP performance after 50 labeled images is only around 17%. Considering that there are few classes with more than 50 images and recalling that the labeled images are counted in the mAP, a good active learning strategy ought to be able to obtain much higher mAP.

V. CONCLUSION

The main conclusions of this paper are threefold: First, we discuss the availability of large cultural heritage image collections that are currently being digitized, and which we believe will be a major topic of interest in the content-based indexing community. Second, by carefully looking at how these collections are currently manually indexed, we detailed specific tasks that are out of the scope of current content-based indexing problems, although they are of great interest for the users of these cultural heritage collections. By performing a review of currently available techniques of content-based indexing, and testing a baseline method on the BnF collection,

Fig. 4. Examples of bad (left) and good (right) performing classes for various features.

we show that there is still a lot of research to be done to achieve satisfactory results.

The main open questions concerning the design of similarity measures specifically tailored for cultural heritage collections are with respect to the wide range in type and scale of signal properties to be encoded in the signature. In particular, it is very difficult with the presented techniques to design a numerical representation that can encode both microscopic properties, such as canvas patterns or brush strokes in painting, and macroscopic properties, such as a scene layout or a direction of illumination. Designing a similarity measure that can tackle all these different types and scales of similarities is probably the biggest challenge in the indexing of cultural heritage collections.

The second problem arises from the paradoxical scarcity of the data. While data are massively available, including cultural heritage images, examples of specific categories may not. For example, only a few examples of an antic coin may be available. In such circumstances, methods that require a large amount of data to train their parameters are hindered and may not be able to obtain satisfactory results. Designing methods able to perform well on very precise similarity tasks with only few relevant training examples is the second challenge in the indexing of cultural heritage collections.

AUTHORS

David Picard

David Picard received the M.Sc. in Electrical Engineering in 2005 and the Ph.D. in image and signal processing in 2008. He joined the ETIS laboratory at the ENSEA (France) in 2010 as an associate professor within the MIDI team. His research interests include image processing and machine learning for visual information retrieval, with focus on kernel methods for multimedia indexing and distributed algorithms.

Philippe-Henri Gosselin

Philippe Henri Gosselin received the PhD degree in image and signal processing in 2005 (Cergy, France). After 2 years of post-doctoral positions at the LIP6 Lab. (Paris, France) and at the ETIS Lab. (Cergy, France), he joined the MIDI Team in the ETIS Lab as an assistant professor, and then was promoted to full professor in 2012. His research focuses on machine learning for online multimedia retrieval. He is involved in several international research projects, with applications to image, video and 3D objects databases.

Marie-Claude Gaspard

Marie-Claude Gaspard is an computer science engineer. She joined the BnF IT Department after several years in service companies. She then moved to the National Bibliographic Agency, where she started working in the field of library science. She completed her training with a master degree in Information and Communication Technologies (ICT). She then headed the team in charge of the BnF Image Bank within the Reproduction Department. She is currently an IT project manager under the direct supervision of the Head of Department.

REFERENCES

[1] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1872–1886, 2013.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.

[3] E. J. Crowley and A. Zisserman, "In search of art," in *Workshop on Computer Vision for Art Analysis, ECCV*, 2014.

[4] ——, "The state of the art: Object retrieval in paintings using discriminative regions," in *British Machine Vision Conference*, 2014.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[6] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*. ACM, 2004, pp. 253–262.

[7] C. Harris and M. Stephens, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

[8] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1704–1716, 2012.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] B. B. Le Cun, J. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*. Citeseer, 1990.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 837–842, 1996.

[13] T. Mensink and J. van Gemert, "The rijksmuseum challenge: Museum-centered visual recognition," in *ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.

[14] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005.

[15] R. Negrel, D. Picard, and P.-H. Gosselin, "Web-scale image retrieval using compact tensor aggregation of visual descriptors," *MultiMedia, IEEE*, vol. 20, no. 3, pp. 24–33, 2013.

[16] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 143–156.

[17] D. Picard and P.-H. Gosselin, "Efficient image signatures and similarities using tensor products of local descriptors," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 680–687, 2013.

[18] B. C. Russell, J. Sivic, J. Ponce, and H. Dessales, "Automatic alignment of paintings and photographs depicting a 3d scene," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 545–552.

[19] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[21] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.

[22] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, Dec 2000.

[23] J. R. Smith and S.-F. Chang, "Visualseek: a fully automated content-based image query system," in *Proceedings of the fourth ACM international conference on Multimedia*. ACM, 1997, pp. 87–98.

[24] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.