

Tag Completion based on Belief Theory and Neighbor Voting

Amel Znaidia
CEA, LIST,
Vision & Content Engineering
Laboratory
Gif-sur-Yvette, France
amel.znaidia@cea.fr

Hervé le Borgne
CEA, LIST,
Vision & Content Engineering
Laboratory
Gif-sur-Yvette, France
herve.le-borgne@cea.fr

Céline Hudelot
Ecole Centrale Paris
Applied Mathematics &
Systems Laboratory
Antony, France
celine.hudelot@ecp.fr

ABSTRACT

We address the problem of tag completion for automatic image annotation. Our method consists in two main steps: creating a list of “candidate tags” from the visual neighbors of the untagged image then using them as pieces of evidence to be combined to provide the final list of predicted tags. Both steps introduce a scheme to tackle with imprecision and uncertainty. First, a bag-of-words (BOW) signature is generated for each neighbor using local soft coding. Second, a sum-pooling operation across the BOW of the k nearest neighbors provides the list of “candidate tags”. Finally, we use neighbors as pieces of evidence to be combined according to the Dempster’s rule to predict the more relevant tags. The method is evaluated in the context of image classification and that of tag suggestion. The database used for visual neighbors search contains 1.2 million images extracted from Flickr. Classification is evaluated on the well known Pascal VOC 2007 and MIR Flickr datasets, on which we obtain similar or better results than the state-of-the-art. For tag suggestion, we manually annotated 241 queries. As well, we obtain competitive results on this task.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Indexing methods; H.2.4 [Database Management]: Systems-Multimedia databases

General Terms

Algorithms, Experimentation

Keywords

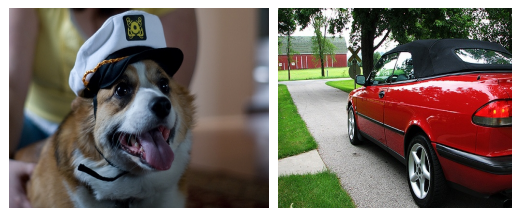
Tag completion, tag suggestion, local soft coding, belief theory, bag of words, image annotation, classification

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMR’13, April 16–20, 2013, Dallas, Texas, USA.

Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...\$15.00.



Tags :
Dog, corgie, 50mm, captain,
Seattle, SonyA200, Minolta.

No Tag

Figure 1: Example of images from Flickr Website with its associated tags.

Online social media services, such as Flickr¹, allow users to share their photos with other people for social interaction. An important feature of these services is that users can annotate their photos with their own tags, in order to facilitate future search and sharing. The main purpose of the users being to make their picture popular to the public, it conflicts with an objective description of the images [1]. Consequently, tags generated by users on Flickr are usually imperfect and only 50% are actually related to the content of the image [12]. These “imperfections” recovers different problems. Above the ones related to the objectivity of users, the human-generated tags (folksonomy) are prone to errors. On the left image of figure 1, only “dog” actually describe the visual content of the image. Other tags are related to shooting conditions (“SonyA200”, “Minolta”, “50mm”) or to subjective context (“captain”, “Seattle”). One could suggest other relevant tags for this image, such as “animal” or “canine”. Other images are simply not annotated at all (figure 1 right), making them almost unusable for research and automatic sharing.

In this paper, we address the problem of tag completion for automatic image annotation. Last trends to generating tags for images without any annotation rely on the idea that if many distinct users use the same tags to label visually similar images, then these tags are likely to reflect the visual contents of the annotated images. Starting from this intuition, classic neighbor voting algorithm use information from the k -nearest neighbors (k NNs) to predict tags [22]. Unfortunately, in the context of social tagging, the tags are freely assigned by users, with various motivations and dif-

¹<http://www.flickr.com/>

ferent judgments on the relevance between a tag and an image. Consequently, tags in social tagging setting are much more uncertain compared to labels in traditional classification problems. In the original voting k NN algorithm, the image is assigned to the majority class according to its k -nearest neighbors, independently of the relevance of each neighbor. Moreover, the classical k NN methods does not deal with ambiguity and imprecise information because of the limitation of the probabilistic framework.

We propose a method to tackle these problems of robustness and effectiveness. First, we start by searching the k nearest neighbors using some visual information. For each neighbor, we compute a bag-of-words (BOW) based signature using locality constraint. Contrary to a classic BoW signature, tags are coded according to several of their closer codewords from a learned codebook. Such a coding has already been proven to be efficient, both for classification from the visual information only [15] and a multimedia context as well [27]. Second, a sum-pooling operation across the BOW of the k nearest neighbors provides the list of “candidate tags”. Finally, basic belief masses are obtained for each nearest neighbor using the distances between this pattern and its neighbors. Their fusion leads to the list of final predicted tags. This last step is derived from the Evidential k NN [4] that apply the Dempster’s rule of combination to a nearest neighbors classifier.

As explained in section 2, our approach differs from existing techniques on two main points. The first novelty is that we use tag corpus knowledge to enrich nearest neighbors description from existing tags. Second and most important difference, we explicitly use a formalism, Belief theory (see section 2.3), which is able to handle neighbors conflict and deal with tag imperfections.

Our work is evaluated in the context of image classification and that of tag suggestion. Classification is evaluated on the well known Pascal VOC 2007 [6] and MIR Flickr [10] datasets, containing 10k to 20k images and 20 to 99 concepts. We evaluated both our method alone as well as lately fused with image-only descriptor for multimodal classification. For tag suggestion, a third database is derived from the one used in [20, 14] for which we created a new ground truth², by manually annotating 241 queries. The database used for visual neighbors search contains 1.2 million images extracted from Flickr.

The remainder of this paper is structured as follows. In Section 2 we describe previous works related to image annotation. A section is specially dedicated to Belief theory and the contextual similarity [18] that is used later to express the similarity between two tags. In Section 3, we introduce our method to tackle with tag completion. Experiments that prove the performance of our approach, in terms of accuracy and robustness, are reported and discussed in Section 4 and 5.

2. RELATED WORK

In this section, we review works closely related to our motivation for tag completion for automatic image annotation. First, we describe previous works related to image annotation. Second, we present the contextual similarity that is

used later to express the similarity between two tags. Finally, we introduce some basic notions of Belief theory.

2.1 Image annotation

Several approaches have been proposed for annotating images by mining the web images with surrounding descriptions. These methods can be classified into two categories: model-based methods and search-based methods.

Model-based methods cast the problem of tag prediction as a binary classification problem where a classifier is learned for each tag. Therefore, almost all classification methods can be applied. One approach is to treat the annotation problem as a translation/projection from images to tags. It is usually performed using the image-tag co-occurrence information [5]. This approach is extended by [17] to latent space using latent semantic analysis techniques. In [13], authors proposed a real-time system based on 2D Multiresolution Hidden Markov Model (MHMM). Images in every category focus on a semantic theme and are described collectively by several words. A category of images is consequently referred to as a semantic concept. Tang *et al.* [21] proposed a k NN-sparse graph-based semi-supervised learning approach for label propagation over noisily-tagged web images.

Unlike model-based approaches, the search-based methods do not need to be constrained in a fixed vocabulary or model. It assumes that images with similar visual content are annotated by similar tags. Recently, nearest neighbor models have been investigated in the annotation community with promising results. Notably, Torralba *et al.* [22], collected about 80 million tiny images, each of which is labeled with one of the 75,062 abstract nouns from WordNet. By fully leveraging on the redundancy of information on the Web, they claimed that with sufficient number of samples, the simple nearest neighbor classifier can achieves reasonable performance for several object/scene detection tasks, when compared with the more sophisticated state-of-the-art techniques. In the same direction, Li *et al.* [14] proposed an algorithm that learns tag relevancy by accumulating votes from visually similar neighbors. In fact, given a user-tagged image, they first perform a k NN search to find its visual neighbors. The tag relevance is determined as the probability that this tag being used to annotate the neighborhood images minus the probability of the tag being used in the entire collection. Wang *et al.* [24] proposed to build a normalized histogram of tags and group names counts from the k -nearest neighbor images. Text classifiers is then trained on the text features. A separate visual classifier is also learned and the final prediction is obtained from a third classifier trained on the confidence values returned by both the textual and the visual classifiers. Guillaumin *et al.* [7] have proposed the tag propagation (TagProp) method to annotate a input image by propagating the tags of the weighted nearest neighbors of that input image. The weighted nearest neighbors were identified by optimally integrating several image similarity metrics. Makadia *et al.* [16] recently have developed the joint equal contribution (JEC) technique, where they used a combination of multiple features and distance metrics to find the nearest neighbors of the input image and a greedy algorithm for transferring tags from visually similar images. Recently, Wu *et al.* [25] proposed a framework for tag completion. They represent the image-tag relation by a tag matrix, and search for the optimal tag matrix consistent

²<http://elm.eeng.dcu.ie/~hlborgne/tagcompletion.html>

with both the observed tags and the pairwise visual similarity between images. This optimization problem is solved using a sub-gradient descent based approach.

Compared with our approach, most existing techniques does not make use of tag corpus knowledge to enrich nearest neighbors description from existing tags. Moreover, there is no explicit use of a formalism which is able to handle neighbors conflict and to deal with tag imperfections.

2.2 Contextual similarity

In this section, we introduce the contextual similarity used later to aggregate image tags into a bag-of-words signature using local soft coding.

In [18], an adaptation of the TF-IDF model to the social space is proposed in order to compute the social relatedness of two tags. Let \mathbf{S} be the matrix of size $N \times K$ defined by:

$$\mathbf{S}(i, j) = \text{users}(\mathbf{t}_i, \mathbf{t}_j) \times \log\left(\frac{\text{users}_{\text{collection}}}{\text{users}_{\text{collection}}(\mathbf{t}_j)}\right), \quad (1)$$

where \mathbf{t}_i is the target tag, \mathbf{t}_j is an element of the codebook, $\text{users}(\mathbf{t}_i, \mathbf{t}_j)$ is the number of distinct users who associate the tag \mathbf{t}_i to the tag \mathbf{t}_j among the top results returned by the Flickr API for \mathbf{t}_i ; $\text{users}_{\text{collection}}(\mathbf{t}_j)$ is the number of distinct users from a pre-fetched subset of Flickr users that have tagged photos with tag \mathbf{t}_j , and N is the number of unique tags associated to photos of the dataset and K is the size of the codebook. Note that some of the tags can have entries on both dimensions of matrix \mathbf{S} . In the current work, we consider a fixed set of tags, that is a tag-codebook.

Relying on this matrix, a Flickr model for a given tag \mathbf{t}_i is proposed in [18] as the following vector of weights:

$$\mathbf{w}_i = [w_{i,1}, w_{i,2}, \dots, w_{i,K}]^T, \quad (2)$$

with $w_{i,j}$ the normalized social weight defined by:

$$w_{i,j} = \frac{\mathbf{S}(i, j)}{\max\{\mathbf{S}(i, k), k = 1, \dots, K\}}. \quad (3)$$

Thereby, given two tag-Flickr models \mathbf{w}_i and \mathbf{w}_j , we compute the contextual similarities between their related tags \mathbf{t}_i and \mathbf{t}_j using the cosine similarity:

$$\text{sim}_{\text{contextual}}(\mathbf{t}_i, \mathbf{t}_j) = \frac{\mathbf{w}_i^T \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|}. \quad (4)$$

2.3 Belief Theory

The belief theory, also called evidence theory or Dempster-Shafer (DS) theory [19] is more and more employed in order to take into account the uncertainties and imprecisions in pattern recognition. The evidence theory is based on the use of functions defined on a *frame of discernment* Ω , represented as the set of all hypothesis in a certain domain. A basic belief assignment (BBA) is a function m that defines the mapping from the power set of Ω to the interval $[0, 1]$ and verifies:

$$m : 2^\Omega \rightarrow [0, 1] \quad (5)$$

$$\sum_{A \in 2^\Omega} m(A) = 1 \quad (6)$$

The quantity $m(A)$ can be interpreted as a measure of the belief that is committed exactly to A , given the available evidence. A subset $A \in 2^\Omega$ with $m(A) > 0$ is called a *focal element* of m .

In DS theory, two functions of evidence can be deduced from m and its associated focal elements, belief function Bel and plausibility function Pl . $Bel(A)$ is the measure of the total belief committed to a set A . The belief function is defined as a mapping $Bel : 2^\Omega \rightarrow [0, 1]$ that satisfies $Bel(\emptyset) = 0, Bel(\Omega) = 1$ and for each focal element A , we have:

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \quad (7)$$

The *plausibility* of A , $Pl(A)$, represents the amounts of belief that could potentially be placed in A and defined as:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \quad (8)$$

When several pieces of evidence are available through their BBA, they can be combined with the Dempster's rule of combination:

$$m_1 \oplus m_2 = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}, & \forall A \subseteq \Omega, A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \quad (9)$$

3. PROPOSED METHOD

The proposed method for tag suggestion using visually similar images is given in figure 2. It consists in two main steps: creating a list of "candidate tags" from the visual neighbors of the untagged image then using them as pieces of evidence to be combined to provide the final list of predicted tags.

Given an untagged image I , we start by searching the k nearest neighbors using visual information (color, texture). First, we compute a BOW signature for each neighbor based on local soft coding. Second, a sum-pooling operation across the BOW of the k nearest neighbors is performed to obtain the list of "candidate tags" (the most frequent). Finally, basic belief masses are obtained for each nearest neighbor using the distances between this pattern and its neighbors. Their fusion leads to the list of final predicted tags.

3.1 Finding candidate tags

Tag-based features bring a complementary description to enrich the semantic description of a given image and we address two issues associated to textual features here. In order to build robust BoW based tag-signatures toward quantization errors, we rely on the locality-constrained coding method that has proved to be effective for visual features when paired with max-pooling [15, 27].

Let I be an untagged image and $\mathcal{N} = \{I^1, \dots, I^k\}$ the set of its nearest neighbors according to a given measure, within an image database. These resources (image database, visual features and similarity function) are not specified at this point but their importance will be discussed later (section 6). Each image I^r has a set of tags $T^r = \{t_1^r \dots t_{n_r}^r\}$. Let consider as well a textual codebook $\mathcal{B} = \{b_1 \dots b_B\}$ that has been built previously (detailed in section 4). Each tag $\mathbf{t}_p^r \in T^r$ is then coded according to its M nearest neighbors of the codebook:

$$z_{p,q} = \begin{cases} \text{sim}_{\text{contextual}}(\mathbf{t}_p^r, \mathbf{b}_q) & \text{if } \mathbf{b}_q \in \mathcal{N}_M(\mathbf{t}_p^r), \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where $\mathcal{N}_M(\mathbf{t}_p^r)$ denotes the M -nearest neighbors of \mathbf{t}_p^r , under the contextual similarity detailed in section 2.2. Another

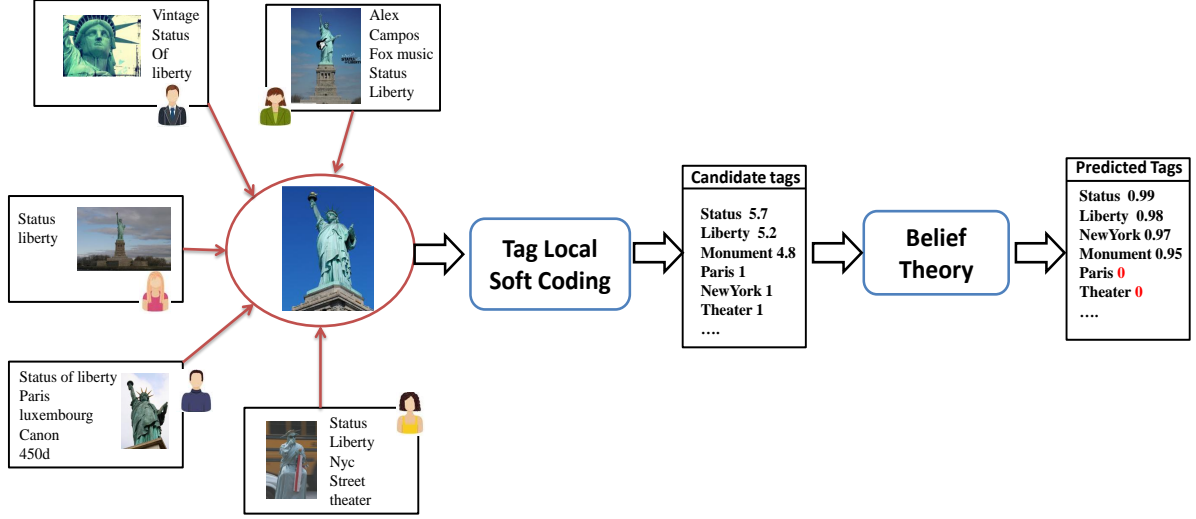


Figure 2: The flowchart of our tag completion approach based on local soft coding and belief theory. First, we compute a bag-of-words (BOW) signature for each neighbor based on local soft coding. Second, a sum-pooling operation across the BOW of the k nearest neighbors is performed to obtain the list of the most frequent tags. Finally, pieces of evidence from neighbors are combined using Dempster’s rule to obtain the set of predicted tags.

similarity between tags and codewords may be used here. The locality assumption in the tag-space induces sparse codes while reducing the reconstruction errors, mainly in terms of semantic reconstruction. The final tag-signature vector results from an aggregation by the maximal values of the coded tags:

$$c_q^r = \max_{p=1}^{Card(T^r)}(z_{p,q}) \quad (11)$$

$\mathbf{c}^r = [c_1^r \dots c_B^r]$ is then the tag-signature vector for the r^{th} neighbor. To obtain the list of “candidate tags”, a sum-pooling operation is performed across the k nearest neighbors tag-signature, as follows:

$$C_q = \sum_{r=1}^k (c_q^r) \quad (12)$$

The entries of $\mathbf{C} = [C_1 \dots C_B]$ with the largest values constitutes the list of “candidate tags”. The number of “candidate tags” kept is arbitrarily set to 10 in our experiments.

3.2 Predicting the final tags

In the following, we denote $\Omega = \{t_1, \dots, t_n\}$ the set of “candidate tags”. Each pair (I^i, t_j) , where $(I^i \in \mathcal{N}, t_j \in \Omega)$, constitutes a distinct item of evidence regarding the relevance of the tag t_j to describe the visual content of the untagged image I . If I is “close” to I^i according to the relevant metric d , then one will be inclined to believe that both images can be tagged with the same tag. On the contrary, if $d(I, I^i)$ is very large, then the consideration of I^i will leave us in a situation of almost complete ignorance concerning the tag t_j . Consequently, this item of evidence may be postulated to induce a Basic Belief Assignment (BBA) $m(\cdot|I^i)$, over the k nearest neighbors, defined by:

$$m(\{t_j\}|I^i) = \alpha\phi_j(d^i) \quad (13)$$

$$m(\Omega|I^i) = 1 - \alpha\phi_j(d^i) \quad (14)$$

where $d^i = d(I, I^i)$ is the distance between the untagged image I and a neighbor I^i , α is a parameter such that $0 < \alpha < 1$ and ϕ_j is a decreasing function verifying $\phi_j(0) = 1$ and $\lim_{d \rightarrow \infty} \phi_j(d) = 0$. One choice for the function ϕ_j can be :

$$\phi_j(d) = \exp(-\gamma_j d^2) \quad (15)$$

In [4], it was proposed to set $\alpha = 0.95$ and γ_j to the inverse of the mean distance between images tagged with the tag t_j . This heuristic yields good results on average. These parameters can be determined also by optimizing a performance criterion as shown in [4]. For simplicity, we choose the first alternative. As a result of considering k nearest neighbors we obtain k BBA for each tag that can be combined using Dempster’s rule of combination to form a final BBA for each tag contained in this neighborhood as follows:

$$m = m(\cdot|I^1) \oplus \dots \oplus m(\cdot|I^k) \quad (16)$$

Adapting this definition, m can be shown to have the following expression:

$$m(\{t_j\}) = \frac{1}{K} (1 - \prod_{i \in \mathcal{N}_j} (1 - \alpha\phi_j(d^i))) \prod_{l \neq j} \prod_{i \in \mathcal{N}_l} (1 - \alpha\phi_l(d^i)) \quad (17)$$

$$m(\Omega) = \frac{1}{K} \prod_{l=1}^n \prod_{i \in \mathcal{N}_l} (1 - \alpha\phi_l(d^i)) \quad (18)$$

where \mathcal{N}_j is the subset of neighbors from \mathcal{N} tagged with the tag t_j and K is a normalization factor. Hence the focal elements of m are singletons and the whole frame Ω . Consequently, the credibility and the plausibility can be defined as follows:

$$bel(\{t_j\}) = m(\{t_j\}) \quad (19)$$

$$pl(\{t_j\}) = m(\{t_j\}) + m(\Omega) \quad (20)$$

These two functions can be used to decide if the tag is relevant to describe the image content. In our case, for fair comparison with the state-of-the-art, we choose to sort the list of “candidate tags” by decreasing credibility values and keep only the p tags with highest values.

4. EXPERIMENTS

The proposed method is evaluated in the context of two applications: image classification and tag suggestion.

4.1 Dataset

For image classification, we report results based on two widely used datasets: MIR Flickr [10] and Pascal VOC 2007 [6]. Both image sets were collected from Flickr but they differ significantly. For instance, MIR Flickr contains a wider variety of concepts whereas Pascal VOC 2007 classes are better balanced.

- **PASCAL VOC 2007** dataset [6] consists of 9,963 images (5011 for training) annotated according to 20 classes. The dataset is one of the most challenging because of the large variation on view size, illumination, scale, deformation and clutter, as well as complex backgrounds. About 38% images are not tagged at all (Table 1).
- **MIR Flickr** dataset [10] consists of 8,000 images for training and 10,000 for testing belonging to 99 highly diversified concepts. These concepts describe the scene (“indoor, outdoor, landscape, etc.”), depicted objects (“car, animal, person etc.”), the type of image content (“portrait, graffiti, art, etc.”), events (“travel, work, etc.”), quality issues (“overexposed, underexposed, blurry, etc.”) and emotions (“funny, cute, nice, scary, etc.”). About 10% images are not tagged at all (Table 1).
- **Flickr 1.2 million** consists of 1.2 million images downloaded from Flickr having no overlap with the untagged images used for test. This collection is used for visual neighbors searching.

For tag suggestion task, we want to evaluate our method on the dataset used in [20, 14]. It consists of 331 images downloaded from Flickr. This dataset is created by manually assessing the relevance of user’s tags with respect to images. An example of images, given in figure 3, shows that the proposed ground truth do not reflect perfectly the images visual content and thus former evaluation of some systems leaded to quite poor results (*e.g.* [14] obtained below 0.15

Table 1: Number and proportion of untagged images in training and test sets, for Pascal VOC 2007 and MIR Flickr datasets.

Dataset	# untagged Train	# untagged Test
Pascal VOC 2007 (prop. total)	1917 (38.3%)	1847 (37.3%)
MIR Flickr (prop. total)	812 (10.1%)	930 (9.3%)



Initial ground truth	2005, february	2006, costume october	2006, Asia, chinese, City travel	2006
Our ground truth	Music, concert Live, Show, Lights, night	People, portrait Makeup, Girl	Baby, sleeping, Bicycle, man Market, Asia	Girl, music Party, Food

Figure 3: Example of images from the dataset of [20]. First row represents ground truth proposed by [20] and the second row represents our manually annotations used as ground truth for tag suggestion evaluation.

MAP and 0.1 Precision@5). We thus decided to manually re-annotate the dataset to better reflect the image visual content. For this, we followed a protocol inspired from the collaborative annotation tool of TrecVid [2] showing that annotating a small fraction of carefully chosen samples of a collection is enough to achieve similar performance (or even better) compared to those obtained with the entire collection. We thus downloaded all the images available on Flickr among the 331, resulting into a collection of 241 images. We run our method as well as two recent ones [24, 14] on these queries to collect potential tags. Then, we manually annotated the queries by keeping the tags that reflect the image visual content³.

4.2 Experimental protocol

We compared our method with two approaches: the Tag Frequency [24] and the Tag Relevancy [14].

- **Tag Frequency** [24]: for a query image, we find its k nearest neighbor images from the auxiliary dataset using visual features. Tags and group name associated with these nearest neighbors are treated as an individual item in the text representation. The text feature is a normalized histogram of tag and group name counts from the k nearest neighbor images.
- **Tag Relevancy** [14]: is calculated by accumulating votes from visually similar neighbors. In fact, given a user-tagged image, they first perform a k NN search to find its visual neighbors. The tag relevance is determined as the probability that this tag being used to annotate the neighborhood images minus the probability of the tag being used in the entire collection.

For the sake of fair comparison, the same processing chain is considered, following literature settings to ensure consistency.

³this collection is available at: <http://elm.eeng.dcu.ie/~hborgne/tagcompletion.html>

Searching Visual Neighbors: The visual similarity between two images are measured by the similarity between their corresponding visual features. Though numerous works have been done for visual feature representation, it is still a challenging problem for content-based image retrieval [22]. For fair comparison, the set of k nearest neighbors used as a starting point of our method is determined according to the visual similarity computed between the same visual feature as [14]. It consists in a combined 64-dimensional global feature for its empirically success in searching millions of web images [23]. It is composed of a 44-dimensional color correlogram in the 44-bin HSV color space [9], 14-dimensional color texture moments [26], and 6-dimensional RGB color moments. The three features are normalized to unit length and concatenated into the final 64D feature. The dissimilarity between images are measured using the Euclidean distance between features. To search for visual neighbors, we adopt K-means clustering based indexing methods. First for indexing, the whole dataset is divided into smaller blocks by K-means clustering. Then for a query, we find neighbors within fewer blocks closest to the query. The search space is thus reduced. For both visual feature extraction and neighbors searching, we use the implementation of [14]. We fixed the number of visual neighbors to 100 for both applications (this number is discussed in section 5.1).

BoW-signature: For the PascalVOC textual codebook, we kept the tags that appear at least 8 times, leading to a dictionary of size 804. In the case of MIR Flickr, we kept the tags used by at least 3 different users, resulting into a textual codebook of 2500 tags. For local soft coding, the neighborhood in the tag feature space was set to 50.

Tag suggestion experiment: for each method, we select the top 5 tags as final suggestion for each untagged image. For tag suggestion, we evaluate directly the performance on these tags.

Image classification experiment: we build a BOW based signature as explained above. A one-versus-all linear kernel based Support Vector Machine (SVM) classifier is learned for each method and we compare their performances in terms of Mean Average Precision (MAP). A separate visual classifier is also learned and fused with the textual classifier learned from each method for a multimodal classification. This visual classifier is totally independent from the one used to search for visual neighbors for untagged images. For the visual classifier, SIFT descriptors are extracted and coded using local soft coding. The patch-size is fixed to 16×16 pixels and the step size for dense sampling to 6 pixels. For the local soft coding, we consider a neighborhood of size 5 and the softness parameter β is set to 10. We use a visual codebook of size 4,000 created using the K-means clustering method on a randomly selected subset of SIFTs from the training dataset ($\approx 10^5$ SIFTs). To aggregate the obtained codes, we perform a max-pooling operation. As well, a spatial pyramid decomposition into 3 levels ($1 \times 1, 2 \times 2, 3 \times 3$) is adopted. Textual and visual classifiers are combined by averaging their corresponding predictions.

5. RESULTS

Before presenting results to both targeted applications (image classification and tag suggestion), we present one experiment to study how results may vary according to the number of visual neighbors considered.

5.1 Impact of visual neighborhood size

The number of nearest neighbors is an important parameter in tag suggestion methods based on nearest neighbors. To analyze the impact of neighborhood size, we tried various values of $k \in \{50, 100, 200, 500\}$ on the Pascal VOC 2007. As shown in figure 4, our method outperforms the

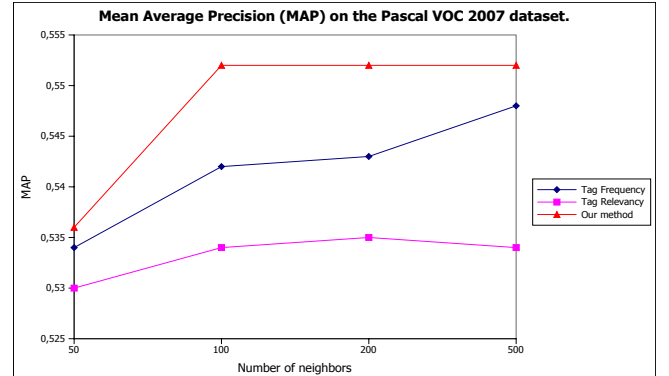


Figure 4: Performance on the Pascal VOC 2007 dataset in terms of Mean Average Precision with respect to the number of nearest neighbors.

two baseline methods for all neighborhood size. Both baseline methods tend to suggest tags occurring frequently in the neighborhood and treat all neighbors equally while the distances between the image and its neighbors are ignored. By contrast, our method starts by predicting new tags by the local soft coding step to enrich the neighbors description and uses the distances to promote the most closed neighbors. In fact, our method reaches the best score (0.552 MAP) with only 100 neighbors and remains stable while varying the neighborhood size. Hence, our method is more effective and stable.

5.2 Tag Suggestion

In Table 2, we report the precision at rank 5 (P@5) on the manually annotated 241 queries. Precision at rank k is defined as the proportion of suggested tags that is relevant, averaged over all photos. As well, we obtain competitive results in tag suggestion task. The tag frequency [24] results are surprisingly better than those of tag relevancy [14] on average. It can be explained by the accuracy of visual search which is query-dependent.

Table 2: Comparison of our system to the state-of-the-art methods on the tag suggestion task.

Method	Average Precision@5
Tag Relevancy [14]	0,349
Tag Frequency [24]	0,387
Our method	0,413

An example of images with suggested tags by the three methods is illustrated in figure 5. As we can see, original tags are imperfect and most of them are subjective. Let's note that these tags are not used in the three methods. Obviously, we can observe that tags predicted by our method are more relevant than those predicted by the two baseline methods. Our approach is more likely to rank relevant tags ahead of

irrelevant ones (shown in bold in figure 5) which is not the case for both tag relevancy and tag frequency.




			
Original tags	Cape cod Bass river lighthouse	plants nature cornwall stonehenge 2005.05.03 xato Vwhiz philip anderson	Iceland Reykjavik 2000.09.03
Tag Relevancy	architecture house tower flag building	food flower salad strawberry red	boat blue city travel boats
Tag Frequency	architecture water house street car	food flower red nature salad	blue street city canon sky
Our Method	architecture house houses blue sky	flowers flower nature red food	blue boat sky Cloud street

Figure 5: Examples of tag suggestion by different methods. The bold font indicates irrelevant suggested tags. Original tags are not used.

5.3 Image classification

In Table 3, we compare the results of the textual classifier based on suggested tags for the three methods on the Pascal VOC 2007 dataset. Results on the MIR Flickr dataset are shown in Table 4. By comparing results on both datasets, we can see that the proposed method based on local soft coding and belief theory gives better results than the baseline methods based on only tag frequency. Our method leads to scores that are three (resp. one) points above the baseline methods on MIR Flickr (resp. Pascal VOC 2007) dataset.

Table 3: Classification performances on PASCAL VOC 07 in terms of Mean Average Precision (MAP), for different methods: 1) based on textual classifier only, 2) based on the combination of textual and visual classifiers

Method	Textual	Multimodal
Tag Relevancy [14]	0.534	0.668
Tag Frequency [24]	0.542	0.676
Our method	0.552	0.684

To demonstrate the effectiveness of suggested tags in multi-modal image classification, a separate visual classifier is also learned and fused with the textual classifier learned from each method as detailed in section 4.2. From results shown in Table 3 and Table 4, we can conclude that our method is more effective than the baseline methods. Furthermore, with a simple combination of linear SVM based output classifiers, we obtain similar or better results than the state-of-the-art on multi-modal image classification on both datasets [8, 3, 27]. These results demonstrate the usefulness of the evidence

formalism to handle neighbors conflict and to deal with tag imperfections. Over the two datasets, our method clearly dominates the remaining methods.

Table 4: Classification performances on MIR Flickr in terms of Mean Average Precision (MAP), for different methods based on: 1) textual classifier only, 2) the combination of textual and visual classifiers.

Method	Textual	Multimodal
Tag Relevancy [14]	0.337	0.412
Tag Frequency [24]	0.343	0.417
Our method	0.37	0.440

In fact, in the baseline methods, the image is assigned to the tag with the majority votes according to its nearest neighbors, independently of the relevance of each neighbor. When nearest neighbors have been tagged subjectively by users, noisy tags will be inevitably assigned to the untagged image due to conflicts or lack of knowledge. First, in the local soft coding step, our method gives a degree of confidence about each tag. Second by exploiting the distance between the untagged image and its nearest neighbors based on belief theory, we are able to reduce the risk of assigning wrongly some tags to an image when the degrees of confidence are not high. That explains the good performances of our method.

6. CONCLUDING REMARKS

We introduced a novel approach for tag suggestion based on local soft coding and belief theory. First, a list of “candidate tags” is created from the visual neighbors of the untagged image, using both local soft coding and two consecutive pooling steps. Then, these tag-signatures are used as pieces of evidence to be combined to provide the final list of predicted tags. This fusion is based on the Dempster’s rule of combination, in accordance with the Evidential k NN framework. Hence, both steps support a scheme to tackle with imprecision and uncertainty that are inherent to this type of information in a social media context. The experiments that we carried out for image classification on two publicly available datasets show that we obtain comparable or better results than the state-of-the-art methods: on Pascal VOC 2007 results are improved of one point both for multimedia and textual-only descriptions; on MIR Flickr, our method leads to scores that are two points above recent state-of-the-art methods. For tag suggestion, we manually annotated 241 queries to propose a new benchmark to the community. For that application as well, we obtained competitive results, with a score two points better than the best recent state-of-the-art method.

Visual neighbors were obtained from an image database containing 1.2 million images extracted from Flickr. This resource is crucial to obtain good raw results for the application considered. Even if our method obtain better results than other recent ones, all of them would benefit from an improved resource. A first direction to improve it is to use a potentially better visual signature to get the neighbors. However, we must keep a certain efficiency in practice to avoid prohibitive time responses to find neighbors. For this, we may search them into a compressed domain that allows to fit large databases into memory [11]. A more difficult

direction of research will be the improvement of the annotation of the resource itself. As we explained, lots of the current annotations are far from being perfect (it is one of the reason we re-annotated the queries to evaluate the work). Hence, this work can naturally be continued into the process of cleaning large multimedia resources.

7. ACKNOWLEDGMENTS

This work is supported by grants from DIGITEO and Région Ile-de-France. We acknowledge support from the French ANR (Agence Nationale de la Recherche) via the PERIPLUS (ANR-10-CORD-026) project, and the Caisse des Dépôts via the EGONOMY project (O12709-67155). We thank Börkur Sigurbjörnsson and Xirong Li for helping us to get the 1.2 million images resource as well as the 331 query identifiers. We are grateful to Xirong Li for the public implementation of their Tag relevance learning algorithm⁴.

8. REFERENCES

- [1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI*, pages 971–980, New York, NY, USA, 2007. ACM.
- [2] S. Ayache and G. Quénot. Evaluation of active learning strategies for video indexing. *Journal of Image Communication*, 22(7-8):692–704, Aug. 2007.
- [3] A. Binder, W. Samek, M. Kloft, C. Müller, K.-R. Müller, and M. Kawanabe. The Joint Submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.
- [4] T. Denoeux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Transaction on systems, man and cybernetics*, 25:804–813, 1995.
- [5] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV '02*, pages 97–112, London, UK, UK, 2002. Springer-Verlag.
- [6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.
- [7] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. *ICCV'09*, pages 309 – 316, Kyoto, Japon, Sept. 2009. IEEE Computer society.
- [8] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. *CVPR '10*, pages 902 – 909, 2010.
- [9] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. *CVPR '97*, Washington, DC, USA, 1997. IEEE Computer Society.
- [10] M. J. Huiskes and M. S. Lew. The MIR flickr retrieval evaluation. In *ACM international conference on Multimedia information retrieval (ICMR)*, pages 39–43, 2008.
- [11] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Sept. 2012.
- [12] L. S. Kennedy, S. fu Chang, and I. V. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. *MIR '06*, pages 249–258, 2006.
- [13] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):985–1002, June 2008.
- [14] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, November 2009.
- [15] L. Liu, L. Wang, and X. Liu. In Defense of Soft-assignment Coding. *ICCV '11*, 2011.
- [16] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *International Journal of Computer Vision*, 90(1):88–105, 2010.
- [17] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. *Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, New York, NY, USA, 2003. ACM.
- [18] A. Popescu and G. Grefenstette. Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 33:1–33:8, 2011.
- [19] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.
- [20] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. *WWW '08*, pages 327–336, New York, NY, USA, 2008. ACM.
- [21] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):14, 2011.
- [22] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, Nov. 2008.
- [23] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Scalable search-based image annotation. *Multimedia Syst.*, 14(4):205–220, 2008.
- [24] G. Wang, D. Hoiem, and D. A. Forsyth. Building text features for object image classification. In *CVPR*, pages 1367–1374, 2009.
- [25] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(Preliminary), 2012.
- [26] H. Yu, M. Li, H.-J. Zhang, and J. Feng. Color texture moments for content-based image retrieval. *ICIP '10*, pages 24–28, 2003.
- [27] A. Znaidia, A. Shabou, A. Popescu, H. Le Borgne, and C. Hudelot. Multimodal feature generation framework for semantic image classification. In *ICMR, International Conference on Multimedia Retrieval, ICMR '12*, Hong Kong, China, June 5-8, 2012, page 38, 2012.

⁴<http://staff.science.uva.nl/~xirong/software/tagrel/index.html>