A Family of Contextual Measures of Similarity between Distributions with Application to Image Retrieval

Florent Perronnin, Yan Liu*and Jean-Michel Renders
Textual and Visual Pattern Analysis (TVPA)
Xerox Research Centre Europe (XRCE), France
{Florent.Perronnin, Yan.Liu, Jean-Michel.Renders}@xerox.com

Abstract

We introduce a novel family of contextual measures of similarity between distributions: the similarity between two distributions q and p is measured in the context of a third distribution u. In our framework any traditional measure of similarity / dissimilarity has its contextual counterpart. We show that for two important families of divergences (Bregman and Csiszár), the contextual similarity computation consists in solving a convex optimization problem. We focus on the case of multinomials and explain how to compute in practice the similarity for several well-known measures.

*These contextual measures are then applied to the image retrieval problem. In such a case, the context u is estimated from the neighbors of a query q. One of the main benefits of our approach lies in the fact that using different contexts, and especially contexts at multiple scales (*i.e.* broad and narrow contexts), provides different views on the same problem. Combining the different views can improve retrieval accuracy. We will show on two very different datasets (one of photographs, the other of document images) that the proposed measures have a relatively small positive impact on macro Average Precision (which measures purely ranking) and a large positive impact on micro Average Precision (which measures both ranking and consistency of the scores across multiple queries).*

1. Introduction

We are interested in image retrieval, the problem which consists in searching in a possibly large dataset of images those templates which are similar to a given query (see *e.g.* [17, 10, 11, 7, 11, 2, 12, 6] for recent works on the topic). When humans have to judge the similarity between two images, and more generally between two objects, they always do so in a given *context*, *i.e.* they do not only consider the two objects to be compared. For instance, while the images of a Maine coon and an American bobtail (two breeds of cats) might be considered similar in the general context of animals, or even in the more focused context of mammals, they can be considered dissimilar in the narrow context of cats. In this simple example, the different contexts correspond to the different scales at which one can consider the problem. We note that different contexts can also correspond to different taxonomies: one can compare two paintings in the context of scenery paintings (semantic context) or in the context of impressionist paintings (artistic context). Different contexts might provide different views on the same problem and different cues may be taken into account to judge the similarity in the different contexts.

There is a significant body of work on contextual (also often referred to as *perceptual*) measures. In the following, we only consider the case where the measure is learned in an unsupervised manner. Our goal is not to provide a full review of the literature on the topic but to give an idea of the variety of approaches which have been proposed.

Several methods consist in modifying not the measure itself but the space in which the measure is computed. This includes Isomap [18], Local Linear Embedding (LLE) [14] or Laplacian Eigenmaps [1]. In [7], Jégou *et al*. propose a contextual measure of distance between points which consists in symmetrizing the K-NN relationship. The initial distance is contextualized by adding a multiplicative penalty term which can be computed iteratively. In effect, it downweights those images which are located in a dense region of the image space. In [20], Zhao *et al*. propose a contextual distance between data points which is defined as the difference of their contributions to the integrity of the structure of the contextual set (defined as the neighboring points).

Closest to our work are those measures typically used in the field of information retrieval. In text retrieval, where a document can be represented as a bag-of-words [16], *i.e.* a histogram of word counts, the term-frequency inverse-document-frequency (TF-IDF) weights the contribution of the different words according to their frequency in a context

[*] Yan Liu is a Ph.D. student in the Laboratoire d'Informatique en Image et Systmes d'Information (LIRIS) at the Ecole Centrale de Lyon (ECL).

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

[15]. The same framework was subsequently applied to image retrieval where images can be described as histograms of visual word counts [17, 10, 7, 11, 2, 12, 6]. The TF-IDF scheme is particularly well-suited to the text retrieval problem where documents are represented as sparse multinomials. However, it cannot be extended in a straightforward manner to dense multinomials or to other distributions than multinomials.

In [13], Ponte and Croft proposed the so-called language modeling (LM) approach to information retrieval. If p, q and u are respectively a template, a query and a context multinomial of dimension D, one can measure the dissimilarity of q and p in the context of u as the Kullback-Leibler (KL) divergence between q and a *smoothed* version of p:

$$\sum_{i=1}^{D} q_i \log \left(\frac{q_i}{\omega p_i + (1-\omega) u_i} \right). \tag{1}$$

Smoothing has two benefits over a standard KL divergence between q and p. First, it avoids $\log(0)$ effects in the case of sparse vectors. Second, by rewriting equation (1) as:

$$-\sum_{i=1}^{D} q_i \log \left(1 + \frac{\omega}{1-\omega} \frac{p_i}{u_i} \right) + C \tag{2}$$

where C is independent of p, we can see that it downweights the influence of frequent words (indices i with large values u_i) as is the case of TF-IDF. A major issue is the sensitivity to the choice of ω (see *e.g.* [19] for a study of the impact of ω as well as different smoothing schemes).

In this article, we introduce a novel family of contextual measures of similarity between distributions. In section 2 we give the definition of our contextual measure and discuss its properties. In our framework any traditional measure of similarity / dissimilarity has its contextual counterpart. We show that when the measure to be contextualized belongs to one of two important families of divergences (Bregman and Csiszár), the contextual similarity computation consists in solving a convex optimization problem. In section 3 we focus on the case of multinomials and show how to compute in practice the contextual similarity for traditional measures. In section 4 we explain how to speed-up retrieval in the case where the optimization scheme is a costly one. In section 5 we apply the contextual measure to the problem of image retrieval and introduce a multi-scale retrieval algorithm. We finally provide in section 6 results on two very different datasets: one of photographs, the other of document images. We will show that the proposed contextual measures have a relatively small positive impact on macro Average Precision (macro-AP) which measures purely ranking and a large positive impact on the micro Average Precision (micro-AP) which measures both ranking and consistency of the scores across multiple queries.

2. Contextual Similarity

We first introduce a broad definition of contextual similarities which is valid for discrete or continuous distributions, parametric or non-parametric distributions, etc.

2.1. Definition

Let p and q be two distributions to be compared and let u be the distribution that models the context. Let f be a "traditional" (i.e. non-contextual) measure of similarity between distributions. We introduce the following function:

$$\phi_f(\omega; q, p, u) = f\left(q, \omega p + (1-\omega) u\right). \tag{3}$$

As we are dealing with distributions, ϕ_f is defined over the interval $0 \leq \omega \leq 1$. We note that in the case where $f(q,p) = E_q[\log p]$, where E_q denotes the expectation under q, $\phi_f(\omega; q, p, u)$ is the distance used in the LM approach to retrieval [13] (c.f. the introduction).

We define the contextual similarity cs_f as:

$$cs_f(q, p|u) = \arg \max_{0 \leq \omega \leq 1} \phi_f(\omega; q, p, u). \tag{4}$$

cs_f is ill-defined for $p = u$ and we choose the convention $cs_f = 1/2$ in such a case.

The intuition behind this measure of similarity is the following one. By maximizing $\phi_f(\omega; q, p, u)$ over ω, we estimate the mixture of p and u that best approximates q. The weight ω which maximizes $\phi_f(\omega; q, p, u)$ reflects how much p contributes to the approximation, i.e. whether q is best modeled by the broad domain information contained in u of the specialized information contained in p. Our similarity is fundamentally different from the traditional LM approach. Especially, there is no parameter tuning required.

By definition cs_f is guaranteed to have values in the interval $[0, 1]$. We note that $q = p \Rightarrow cs_f(\omega; q, p, u) = 1$ but that the converse does not hold. ϕ_f and thus cs_f are asymmetric in p and q even if f is symmetric, i.e. $cs_f(q, p|u) \neq cs_f(p, q|u)$ in general. There exist various ways to symmetrize the contextual similarity if needed. One way is to combine $cs_f(p, q|u)$ and $cs_f(q, p|u)$ using for instance a sum or product rule. Another way is to symmetrize ϕ_f, e.g. as follows:

$$\begin{aligned} \phi_f(\omega; q, p, u) &= f\left(q, \omega p + (1-\omega) u\right) \\ &+ f\left(p, \omega q + (1-\omega) u\right). \end{aligned} \tag{5}$$

In our experiments, we always made us of the symmetric contextual measure.

2.2. Choice of the function f

We have not yet defined a similarity measure but a family of similarity measures parametrized by the particular choice of the function f. cs_f can thus be understood as a contextualized version of f. f can be virtually any measure of similarity between distributions. Obviously, f can be a dissimilarity instead of a similarity: this just requires changing the $\max$ by a $\min$ in (4).

Interestingly, not all measures f are good candidates for contextualization. A simple counter-example is the Expected Likelihood (EL) kernel [5]: $EL(q,p) = E_q[p] = E_p[q]$ (in the case of multinomial distributions, this is simply the dot-product). Except in the case where $E_q[p-u] = 0$, is is easy to show that cs_{EL} gives binary values (0/1).

It is advantageous to choose ϕ to be concave (resp. convex) in ω if f is a similarity (resp. dissimilarity) as one is thus guaranteed to have a unique optimum which simplifies the optimization process. In the following, we consider the case of discrete finite distributions (such as multinomials). We show that when f belongs to one of two important families of divergences, ϕ_f is convex in ω.

Bregman divergences. The Bregman divergence between two distributions x and y (x and y belong to the space of probabilities Ω) for a convex function $h : \Omega \rightarrow \mathbb{R}$ is defined as:

$$B_h(x,y) = h(x) - h(y) - \langle \nabla h(y), (x-y) \rangle \tag{6}$$

where ∇h denotes the gradient vector of h and $\langle .,. \rangle$ the dot product. Intuitively, $B_h(x,y)$ can be understood as the difference between the value of h at point x and the value of the first-order Taylor expansion of h around y evaluated at x. Special cases of Bregman divergences include the Euclidean distance, the Mahalanobis distance, the Kullback-Leibler divergence or the Itakura-Saito divergence.

If $\phi(\omega; q, p, u) = B_h(\omega p + (1-\omega)u, q)$, then $\phi(\omega; q, p, u)$ is convex in ω. To prove this assertion, it is sufficient to show that the second order derivative is positive. We have:

$$\begin{aligned} &\frac{\partial^2}{\partial \omega^2} B_h(\omega p + (1-\omega)u, q) \\ &= (p-u)^T \nabla^2 h(\omega p + (1-\omega)u)(p-u) \end{aligned} \tag{7}$$

where $\nabla^2 h$ denotes the Hessian matrix of h and ^{T} the transposition. As h is convex, this quantity is positive by definition and thus ϕ is convex in ω.

We note however that if $\phi(\omega; q, p, u) = B_h(q, \omega p + (1-\omega)u)$, we cannot conclude on the convexity of ϕ (the second order derivative with respect to ω includes third order derivatives of h).

Csiszár divergences. The Csiszár divergence between two discrete distributions x and y for a convex function $h : \mathbb{R} \rightarrow \mathbb{R}$ is given by:

$$f_h(x,y) = \sum_i x_i h\left(\frac{y_i}{x_i}\right). \tag{8}$$

Special cases of Csiszár divergences include the Manhattan distance, the Kullback-Leibler divergence, the Hellinger distance or the Rényi divergence.

If $\phi(\omega; q, p, u) = f_h(q, \omega p + (1-\omega)u)$, then $\phi(\omega; q, p, u)$ is convex in ω. One more time, it is sufficient to show that the second order derivative is positive. We have:

$$\begin{aligned} &\frac{\partial^2}{\partial \omega^2} f_h(q, \omega p + (1-\omega)u) \\ &= \sum_i \frac{(p_i - u_i)^2}{q_i} h''\left(\frac{\omega p_i + (1-\omega)u_i}{q_i}\right) \end{aligned} \tag{9}$$

where h'' is the second order derivative of h. As h is convex, $h'' \geq 0$ and the previous quantity is positive.

Similarly, if $\phi(\omega; q, p, u) = f_h(\omega p + (1-\omega)u, q)$, ϕ is convex in ω as:

$$\begin{aligned} &\frac{\partial^2}{\partial \omega^2} f_h(\omega p + (1-\omega)u, q) \\ &= \sum_i \frac{q_i^2 (p_i - u_i)^2}{(\omega p_i + (1-\omega)u_i)^3} h''\left(\frac{q_i}{\omega p_i + (1-\omega)u_i}\right) \end{aligned} \tag{10}$$

is a positive quantity. These results can be easily extended to the case of continuous distributions (replacing the sum by an integral).

3. Multinomial Distributions

We now assume that p, q and u are multinomials of dimensionality D. We will consider typical measures between multinomials and show how to compute their contextual counterparts in practice.

3.1. Euclidean distance (L2)

Taking the derivative of ϕ_{L2} and equating it to zero trivially leads to the following closed-form formula:

$$\omega = \frac{\sum_{i=1}^{D} (p_i - u_i)(q_i - u_i)}{\sum_{i=1}^{D} (p_i - u_i)^2}. \tag{11}$$

As we have to enforce the constraint $0 \leq \omega \leq 1$, if the value computed with (11) is lower than 0 (resp. greater than 1), it is forced to 0 (resp. 1). This measure of similarity has a simple geometric interpretation: cs_{L2} is proportional to the projection of the vector $(q-u)$ on $(p-u)$.

3.2. Manhattan distance (L1)

By definition we have:

$$\phi_{L1}(\omega; q, p, u) = \sum_{i: p_i - u_i \neq 0} |p_i - u_i| \left| \omega - \frac{q_i - u_i}{p_i - u_i} \right| \tag{12}$$

This function is convex and piecewise linear. Thus, its minimum is necessarily reached at one of the values $(q_i - u_i)/(p_i - u_i)$. The minimization of ϕ_{L1} is a weighted median problem which can be solved in $O(D)$.

3.3. Kullback Leibler (KL)

By definition we have:

$$\phi_{KL}(\omega; q, p, u) = \sum_{i=1}^{D} q_i \log \left(\frac{q_i}{\omega p_i + (1-\omega) u_i} \right). \quad (13)$$

This objective function is similar to that of Probabilistic Latent Semantic Analysis (PLSA) [4] and can thus be optimized iteratively using the expectation maximization algorithm. At iteration $(k+1)$, we have:

E-step:

$$\gamma_i^{(k+1)} = \frac{\omega^{(k)} p_i}{\omega^{(k)} p_i + (1-\omega^{(k)}) u_i}. \quad (14)$$

M-step:

$$\omega^{(k+1)} = \sum_{i=1}^{D} q_i \gamma_i^{(k+1)}. \quad (15)$$

We note however that EM is slow to converge. Hence, faster optimization techniques such as gradient-descent type algorithms can be considered.

3.4. Other Measures

The Hellinger (HE) distance is defined as:

$$\sum_{i=1}^{D} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2, \quad (16)$$

and the χ^2 (X2) as:

$$\frac{1}{2} \sum_{i=1}^{D} \frac{(p_i - q_i)^2}{p_i + q_i}. \quad (17)$$

These measures of similarity lead to convex objective functions ϕ. However, there is no closed form formula for cs_{HE} and cs_{X2}. Therefore, we have to resort to gradient-based methods for the optimization.

4. Speeding-up Retrieval

For certain measures f, such as the KL divergence, the Hellinger distance or the χ^2, computing cs_f requires an iterative optimization scheme. This might be too computationally intensive for large-scale retrieval. In the case of retrieval problems, we may not be interested in the exact value of the similarity. We may just want to know whether the similarity exceeds a threshold θ.

We introduce:

$$\psi_f(\theta; q, p, u) = \left. \frac{\partial}{\partial \omega} \right|_{\omega=\theta} \phi_f(\omega; q, p, u). \quad (18)$$

The gradient function $\psi_f(\theta; q, p, u)$ is in itself a family of contextual measures of similarity which is doubly parametrized by the function f and the parameter θ. The main advantage of $\psi_f(\theta; q, p, u)$ over $cs_f(q, p|u)$ is that the former one is faster to estimate. For instance, in the case where f is the KL divergence we get:

$$\psi_{KL}(\theta; q, p, u) = \sum_{i=1}^{D} q_i \frac{p_i - u_i}{\theta p_i + (1-\theta) u_i}. \quad (19)$$

If the values $\frac{(p_i - u_i)}{\theta p_i + (1-\theta) u_i}$ can be pre-computed, this quantity is very efficient to evaluate (dot product). In the case of other measures f, simple closed-form gradient formulae can also be derived. However, in preliminary experiments (not reported in this paper), gradient measures generally led to a lower retrieval accuracy compared the contextual measure cs_f. They also require the tuning of the parameter θ (as is the case of the LM approach to retrieval).

$\psi_f(\theta; q, p, u)$ may be used as a complement to $cs_f(q, p|u)$ to speed-up retrieval. If ϕ if differentiable at the point $\omega = \theta$ and if ϕ is concave in ω, we have the following equivalence:

$$\arg\max_{\omega} \phi_f(\omega; q, p, u) \geq \theta \Leftrightarrow \psi_f(\theta; q, p, u) \geq 0. \quad (20)$$

For a given query q, if one wants to retrieve and rank all the templates p whose similarity $cs_f(q, p|u)$ is above θ then one can use a two-step approach:

1. Compute the cheap gradient similarities $\psi_f(\theta; q, p, u)$ for all templates p.

2. Compute the more costly similarities $cs_f(q, p|u)$ for all templates p such that $\psi_f(\theta; q, p, u) \geq 0$.

The second step is not required if we choose $\theta = 0$ (resp. $\theta = 1$) and we have $\psi_f(0; q, p, u) \leq 0$ (resp. $\psi_f(1; q, p, u) \geq 0$) as in such a case we are guaranteed to have $cs_f(q, p|u) = 0$ (resp. $cs_f(q, p|u) = 1$). This can provide very substantial savings (c.f. next section).

5. Retrieval with Multiple Contexts

We assume that the objects to be retrieved can be modeled by distributions. We have a query q and a set of N templates: $\{p_i, i = 1...N\}$. Our goal is to score and rank all templates. The main issue is the choice of the context model u. As the templates themselves provide a context for the query, u may be estimated using all p_i's. However, this may lead to poor results in the case depicted on Figure 1. We consider a toy example with 3 classes (each class is represented by an ellipse in the space of distributions) where the distance between classes 1 and 2 is significantly smaller than the distance between classes 1 and 3 and 2 and 3. For instance, classes 1, 2 and 3 may correspond to cat, dog and

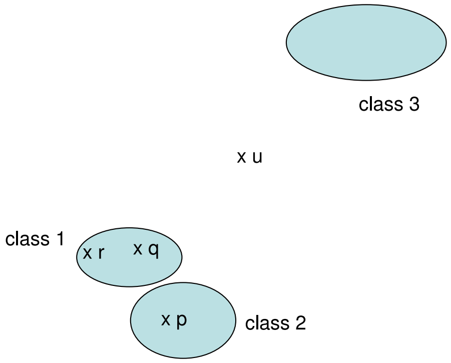

Figure 1. Typical case where a single context estimated on the whole dataset is insufficient for retrieval.

cow images respectively. Ideally, we would like to have the similarity between q and r to be higher than the similarity between q and p for all q and r in class 1 and all p in class 2. As the context u is at a significant distance from class 1 and class 2, we might have $cs_f(q, r|u) \approx cs_f(q, p|u)$ In our animal example, this means that cats and dogs are very similar in the context of cats, dogs and cows.

We can however improve retrieval by using different contexts for different queries. As the choice of the best context for a given query might be a difficult one, we propose to use multiple contexts per query (*i.e.* contexts at multiple scales) and average the similarities across contexts. We will show experimentally in the next section that averaging across multiple scales makes sense because of the complementary information contained in each scale. Broad contexts typically lead to coarse measures of similarity (low precision at low recall but relatively good precision at high recall) while narrow contexts lead to fine measures (high precision at low recall but low precision at high recall).

We thus propose the following algorithm. Given a query q, for scales $k = 1...K$:

1. Compute the similarity $f(q, p_i)$ to all templates p_i and keep N_k-closest, where N_k increases with k. Let $\mathcal{L}_k$ be the list of their indices.

2. Estimate the context u_k as the centroid of the N_k templates. If f is a similarity measure, it is computed as:

$$u_k = \arg\max_u \sum_{i \in \mathcal{L}_k} f(u, p_i). \qquad (21)$$

   The previous optimization has to be done under the constraint that u_k is a multinomial. In the case of L2 or KL, u_k is simply the average.

3. For all templates compute:

$$\omega_{i,k} = cs_f(q, p_i|u_k) \qquad (22)$$

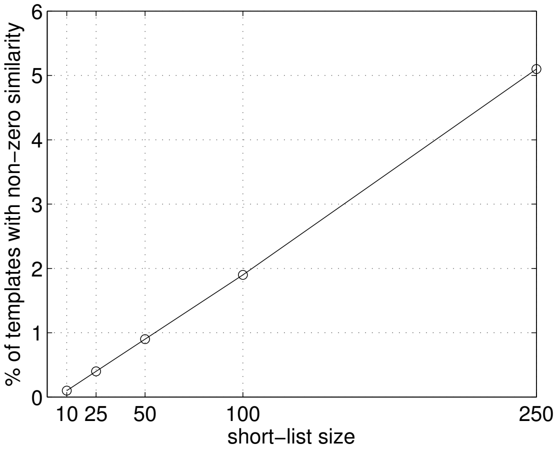

Figure 2. Average percentage of templates which have a non-zero similarity to a given query as a function of the short-list size as estimated on the Holiday dataset [6]. Notice the near-perfect linear dependency.

The multi-scale similarity for template p_i is the average of the $\omega_{i,k}$'s for the different k's. We experimented with two simple averaging schemes. The first one is a the arithmetic mean. The second one is a weighted mean where the weight at a given scale is proportional to $1/N_k$. This gives more weight to fine measures than coarse ones. In our experiments, we chose the sequence of N_k's to increase (approximately) exponentially, *e.g.* 10, 25, 50, 100, etc.

The cost of computing $cs_f(q, p_i|u_k)$ at step 3 of the algorithm might be greatly reduced by using the trick introduced in the previous section. Indeed, the more focused the context, *i.e.* the smaller the number N_k, the smaller the number of templates with non-zero values $cs_f(q, p_i|u_k)$. We show on Figure 2 on the Holiday dataset how this number evolves with N_k (c.f. section 6.2 for more details on the experimental setup). For instance, for $N_k = 10$ on the average on the order of 0.1% of the templates have a non-zero similarity to the query (which corresponds on this dataset to less than 2 templates). We found-out experimentally that the probability for a template p_i to have a non-zero value $cs_f(q, p_i|u_k)$ if p_i is not in the short-list $\mathcal{L}_k$ was close to zero. Therefore, we further speed-up the retrieval of the algorithm by modifying step 3 as follows:

$$\omega_{i,k} = \begin{cases} cs_f(q, p_i|u_k) & \text{if } i \in \mathcal{L}_k \\ 0 & \text{if } i \notin \mathcal{L}_k \end{cases} \qquad (23)$$

with no significant difference in performance.

We note that our multi-scale retrieval algorithm has a flavor of pseudo-relevance feedback (PRF) as we use images which are similar to a query for re-scoring (see *e.g.* [2] for an example of PRF as applied to image retrieval). However, the proposed approach is significantly different for the two following reasons. First, we do not use the similar images to re-estimate the query model but to estimate a context model. Second, PRF typically uses only few images (hopefully only relevant ones) to update the query while we use a larger set of images, most of which might be irrelevant. We will show in section 6.2 that our multi-scale retrieval can significantly improve accuracy on a dataset where PRF would be useless.

6. Experimental Validation

We first discuss measures of retrieval accuracy. We then report results on two very different datasets: the first one of photographs, the second one of document images.

6.1. Measures of retrieval accuracy

Image retrieval – and more generally information retrieval – has traditionally been considered as a pure ranking problem. The assumption is that the system does not know the intent of the user and therefore that it should return all templates (or at least a large subset of them) in descending rank order and let the user choose the relevant ones. For instance, if a user queries a database with a cat image, does this mean that he / she is interested in retrieving all sorts of cats or only cats of the same breed? However, there exist applications where the intent of the user may be known. For instance, one can be interested in retrieving images of the same object [10], of the same architectural landmark [11] or of the same scene [6]. In such a case, the ability to retrieve only relevant images could be of high value to the user.

Therefore, we use two measures of retrieval accuracy:

- Macro Average Precision (macro-AP) consists in computing the AP for each query separately and then averaging these values. Macro-AP only measures ranking performance.

- Micro Average Precision (micro-AP) consists in computing the AP for all queries simultaneously. Micro-AP measures both ranking performance as well as the ability to set a common threshold across different queries.

6.2. Holiday dataset

The Holiday dataset [6] contains 1,491 images of personal holiday photos. There are 500 image groups, each of which represents a distinct scene. The first image of each group is the query and the correct retrieval results are the other images of the group. This dataset is perfect to show that our multi-scale retrieval algorithm is different from PRF. Indeed, it contains on the average two relevant images per query and more than half of the queries have a single relevant image. PRF would thus be useless on such a dataset.

		L2	L1	KL	HE	X2
micro	Base	18.5	16.7	14.5	16.3	16.8
	Ctxt	37.9	47.0	45.1	47.0	46.3
macro	Base	45.7	55.0	57.9	55.3	55.8
	Ctxt	51.9	60.0	59.1	60.4	60.4

Table 1. Results on the holiday dataset in terms of micro- and macro-AP (in %). “Base” = baseline measures. “Ctxt” = proposed measure of contextual similarity.

To encode images, we adopt the bag-of-visual-words (BOV) framework [17, 3]. Low-level features [9] are extracted on dense grids at multiple scales. Offline, we learn a visual vocabulary containing approximately 4,000 visual words through clustering of a large set of low-level features. Following [6], the visual vocabulary is learned on a separate dataset (in our case, a set of images coming from a photofinishing workflow). Each image is then encoded as a histogram of the number of occurrences of each visual word, *i.e*. a multinomial distribution.

Baseline. We first report the results of baseline measures in Table 1. As the KL divergence is not defined in the case of sparse multinomials, the baseline KL results reported are that of the LM approach of [13] which measures the KL between the query and a smoothed version of the template (c.f. introduction). The multinomial u in equation (1) was estimated through averaging of all the BOV histograms in the dataset. We first determined in a set of preliminary experiments the optimal smoothing factor ω. This means that the value ω was tuned to optimize the retrieval accuracy on this dataset, which gives an unfair advantage to the LM approach with respect to other baseline measures or our approach. In the following experiments it is set to $\omega = 0.1$.

Although our goal is not on comparison with [6] (especially [6] only reports macro-AP and not micro-AP), we can see that our best baseline is on par with their baseline BOV in term of macro-AP (54.9% with a 200K words visual vocabulary).

We note that results in terms of micro-AP are significantly lower than that of macro-AP. This shows that with simple measures of similarity, using the same decision threshold across different queries leads to poor results, even when the different queries correspond to the same task.

Contextual KL. We now focus on the proposed contextual KL similarity (c.f. section 3.3). We first consider a single scale, *i.e*. we do not perform the averaging operation over multiple scales. We report results in Figure 3 as a function of the short-list size N_k used to estimate the context (c.f. section 5). Varying the short-list size has a significant impact both on the micro-AP and macro-AP. Note that we did similar experiments with the LM approach to retrieval

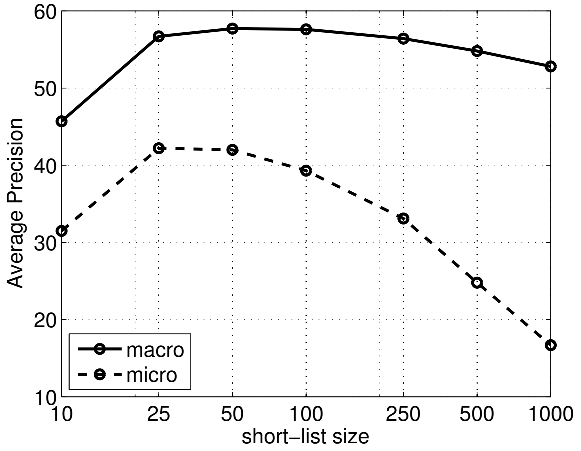

Figure 3. Results (in terms of micro-AP and macro-AP) of the proposed contextual KL for various short-list sizes (single scale).

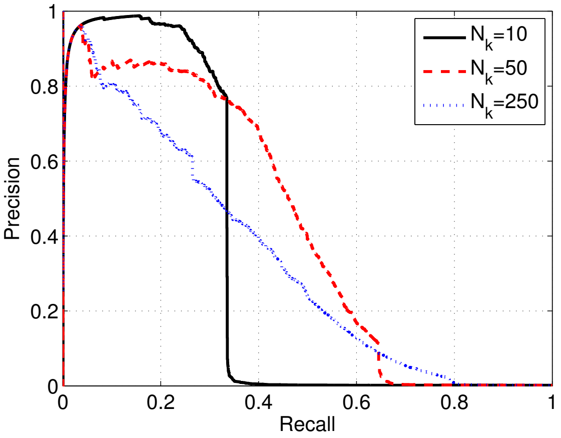

Figure 4. Precision / Recall curves for the contextual KL for various short list-sizes.

but that varying the short-list size had virtually no impact on the micro-AP or macro-AP.

We show on Figure 4 the Precision / Recall curves for the contextual KL for various short-list sizes. Narrow contexts ($N_k = 10$) lead to a high precision at low recall but low precision at higher recalls. The opposite effect can be observed for broad contexts ($N_k = 250$). This clearly shows that different contexts can contain complementary information. We will show that it is beneficial to combine them as this (i) avoids the difficult choice of choosing a priori the best context and (ii) this can lead to a higher retrieval accuracy than each context considered separately.

We now report results when averaging up to a certain scale on Figure 5. We can see that both averaging schemes give similar performance for the macro-AP but that the weighted average is more robust for micro-AP. The best results are 45.1% (micro-AP) and 59.1% (macro-AP).

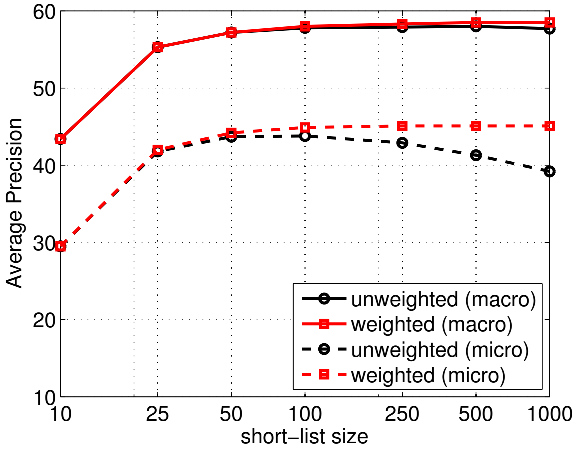

Figure 5. Comparison (in terms of micro-AP and macro-AP) of the unweighted and weighted averages when averaging up to a certain scale for the contextual KL divergence. For instance the results for a short-list size of 50 correspond to the averaging of $N_k = 10$, 25 and 50.

Other measures of similarity. We now consider other measures of similarity to show that our good results are not limited to the KL divergence. In table 1, we report results with our best system, *i.e.* computing the contextual similarity at multiple scales and then computing the weighted average. All contextual measures improve over their non-contextual counterpart in terms of macro-AP (from 1.2% absolute for the KL up to 6.2% for the L2). However, the improvement in terms of micro-AP is much more dramatic (on the order of 30% absolute).

6.3. Document dataset

We now report experimental results on an internal dataset of document images. This dataset contains 1,400 black and white (*i.e.* binary) images. There are 14 classes (100 images per class). This dataset contains both highly structured documents (*e.g.* forms) and loosely structured documents (*e.g.* handwritten letters).

We experimented with two image representations. The first one is the BOV as was the case for the holiday dataset. The second one is based on run-length (RL) histograms [8]. In a nutshell, here is the principle of RL histograms. A run is a sequence of pixels with the same value. The length of a run is the number of pixels such a sequence contains. The RL histogram is a histogram of the lengths of the runs for black pixels and white pixels in 4 directions (horizontal, vertical, diagonal and anti-diagonal). Although the BOV representation led to slightly better results than the RL on this dataset, we prefer RL as it does not require any training phase. Also, providing results on another representation will show that our good results are not limited to the BOV. A difference between RL and BOV histograms is that the lat-

		L2	L1	KL	HE	X2
micro	Base	50.1	59.7	56.5	55.3	57.3
	Ctxt	65.8	72.5	70.0	70.7	70.1
macro	Base	61.5	67.7	65.4	64.6	66.1
	Ctxt	67.1	72.8	70.1	70.9	70.2

Table 2. Results on the document dataset in terms of micro- and macro-AP (in %) "Base" = baseline measures. "Ctxt" = proposed measure of contextual similarity.

ter one is very sparse while the former one is almost dense.

We applied the proposed algorithm with the same settings as in the case of the holiday dataset. Results are shown in Table 2 for baseline measures as well as the proposed contextual measures. One more time, as the KL is not defined when some of the multinomial values are zero, we report results for the smoothed version of the KL. Again, the ω value for the smooth KL was tuned on this dataset. The absolute increase in accuracy for the proposed approach ranges from 4.1% (X2) to 6.3% (HE) for the macro-AP and from 12.8% (L1,X2) to 15.7% (L2) for the micro-AP.

7. Conclusion

In this article, we presented a novel family of contextual measures of similarity between distributions. We explained that in our framework any measure of similarity or dissimilarity had its contextual counterpart. We showed that for two important families of divergences (Bregman and Csiszár) the contextual similarity computation is a convex optimization problem. We focused on the case of multinomials and explained how to compute in practice the similarity for several well-known measures.

This framework was applied to the image retrieval problem. In such a case, the context is estimated from the neighbors of a query. We explained that using multiple contexts (*i.e.* different sizes of neighborhoods) was beneficial as different contexts contain complementary information. Experiments carried out on two very different datasets (the first one of photographs, the second one of document images) showed small consistent improvements in terms of macro-AP (which measures purely ranking) and large improvements in term of micro-AP (which measures both ranking and stability of the scores across multiple queries).

In the future, we intend to focus on the application of this framework to clustering. Indeed, clustering consists in grouping "similar" images where the notion of similarity depends on the other images contained in the dataset. For instance, while it might make sense to group images of different breeds of cats in a general dataset of animal images, it might not in a dataset of cat images. Hence, we believe that clustering is a problem that could benefit greatly from the proposed family of measures.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computing*, 15(6), 2003.

[2] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE ICCV*, 2007.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.

[4] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[5] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *COLT*, pages 57–73, 2003.

[6] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *IEEE ECCV*, 2008.

[7] H. Jégou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *IEEE CVPR*, 2007.

[8] D. Keysers, F. Shafait, and T. Breuel. Document image zone classification - a simple high-performance approach. In *VISAPP*, 2007.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[10] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE CVPR*, 2006.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE CVPR*, 2007.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE CVPR*, 2008.

[13] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *ACM SIGIR*, page 1998.

[14] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2000.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, 1988.

[16] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[17] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE ICCV*, 2003.

[18] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2000.

[19] C. Zhai and J. Lafferty. A study of smoothing methods for language model applied to ad hoc information retrieval. In *ACM SIGIR*, 2001.

[20] D. Zhao, Z. Lin, and X. Tang. Contextual distance for data perception. In *IEEE ICCV*, 2007.