Diversification with Fisher Kernel Pseudo-Feedback

Bogdan Boteanu, Ionuţ Mironică, Bogdan Ionescu,
LAPI, University ”Politehnica” of Bucharest, 061071 Bucharest, Romania,
Email: $\{$*bboteanu, imironica, bionescu*$\}$*@alpha.imag.pub.ro*

Abstract— **In this article we approach the problem of image search result diversification from a novel perspective that involves the use of relevance feedback (RF). Traditional RF introduces the user in the processing loop by harvesting feedback about the relevance of the search results. This information is used for recomputing a better representation of the data needed. The novelty of our approach is twofold. First, we exploit the RF concept in a completely automated manner via pseudo-relevance feedback; this is while addressing the diversification in priority rather than the relevance. Secondly, we introduce a more efficient visual content representation scheme that exploits Fisher Kernels (FK). It allows to better capture variability of visual keypoints information. We use unsupervised hierarchical clustering to regroup FK descriptors in classes. Diversification is finally achieved with a re-ranking approach. Experimental validation on Flickr data shows the advantages of this approach.**

I. Introduction

An efficient retrieval system should be able to *summarize* search results and give a global view so that it surfaces results that are both *relevant* and covering different aspects, i.e., *diverse*, of the query. Relevance was more thoroughly studied in existing literature than diversification [1] and even though a considerable amount of diversification literature exists (mainly in the text-retrieval), the topic remains important, especially in multimedia [2].

The key of the entire diversification process is to mitigate the two components, relevance and diversity, which in general tend to be antinomic: too much diversification may result in losing relevant items while increasing solely the relevance will tend to provide many near duplicates. For instance, the authors in [3] use lightweight clustering in combination with a dynamic weighting function of visual features to best capture the discriminative aspects of image results; or the authors in [2] who address the problem of image diversification in the context of automatic visual summarization of geographic areas and exploits user-contributed images and related explicit and implicit metadata collected from popular content-sharing websites. The approach is based on a Random walk scheme with restarts over a graph that models relations between images, visual features, associated text, as well as the information on the uploader and commentators.

In this paper we approach the diversification problem from a new perspective via relevance feedback. Relevance feedback has proven to increase retrieval accuracy and gives more personalized results for the user. One of the earliest and most successful RF algorithms is the Rocchio's algorithm [4] (which is still used at the present time). Using the set of relevant and non-relevant documents selected from the current user relevance feedback window, the Rocchio's algorithm modifies the features of the initial query by adding the features of positive examples and subtracting the features of negative examples to the original feature. Another relevant approach is the Relevance Feature Estimation (RFE) algorithm [5]. It assumes that for a given query, according to the user's subjective judgment, some specific features may be more important than others. A re-weighting strategy is adopted which analyzes the relevant objects in order to understand which dimensions are more important than others in determining “what makes a result relevant”. Features with higher variance with respect to the relevant queries lead to lower importance factors than elements with reduced variation.

More recently, machine learning techniques found their application in relevance feedback approaches. The relevance feedback problem can be formulated either as a two class classification of the negative and positive samples; or as an one class classification problem, i.e., separate positive samples by negative samples. After a training step, all the results are ranked according to the classifiers's confidence level [6], or classified as relevant or non-relevant depending on some output functions [7]. Some of the most successful techniques use Support Vector Machines [6], Nearest Neighbor approaches [3], classification trees, e.g., use of Random Forests; or boosting techniques, e.g., AdaBoost [7].

Almost all the existing relevance feedback techniques focus exclusively on improving the relevance of the results.

In this paper we propose a novel pseudo-relevance perspective that exploits the concept of relevance feedback while pushing in priority the diversification, in an automated manner. User feedback is simulated automatically by selecting positive and negative examples from the initial query results. Then, we employ a highly efficient description scheme that uses Fisher Kernel representations to capture variability in keypoints information. Unsupervised hierarchical clustering is then used to re-group images according to their contents. Diversification is finally achieved with a re-ranking approach. Experimental validation shows the benefits of this approach allowing to improve the retrieval diversification capabilities.

The reminder of the paper is organized as following. Section II describes the proposed approach. Section III deals with the experimental validation and Section IV concludes the paper and discusses future perspectives.

Part of this work was supported under InnoRESEARCH POSDRU/159/1.5/S/132395 (2014-2015).

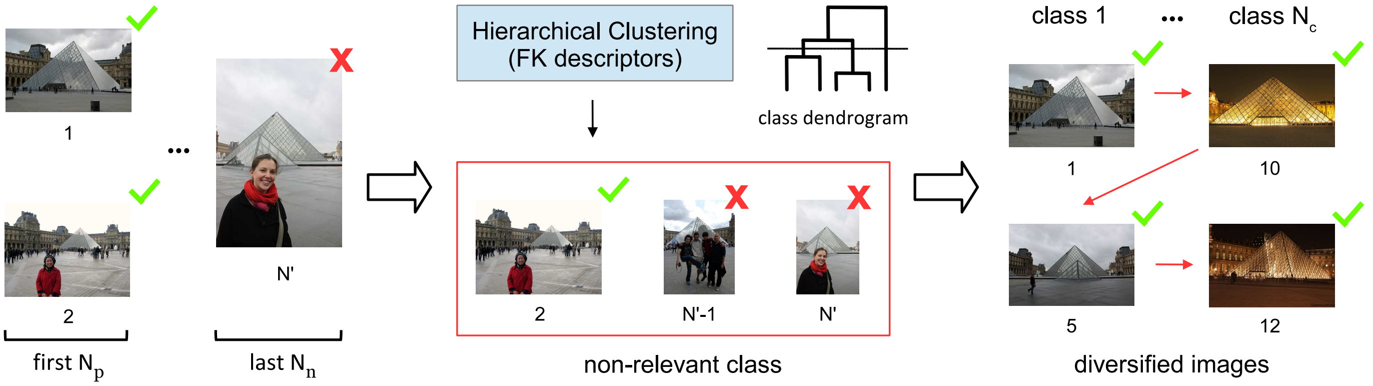

Fig. 1: General scheme of the proposed approach: Selection of positive and negative examples (N_p and N_n, respectively; N' is the total number of returned images), Clustering and pruning (N_c is the number of resulting classes), Diversification.

II. Proposed approach

The proposed approach operates on top of an existing retrieval system and works as a re-ranking step that refines the initial query results. The architecture of the proposed system is illustrated in Figure 1. In the first step, some positive and negative examples are selected from the query results. Then the visual content is represented with a new description scheme that uses Fisher Kernel representation. An unsupervised classification step is then used to cluster the FK descriptors of the selected examples. The final step consists of a cluster diversification strategy. Each of the processing steps is detailed in the following.

A. *Selection of positive and negative examples*

Instead of using a classic relevance feedback strategy where the user is supposed to provide positive examples, we use a pseudo-relevance feedback assumption [9]. We retain the first N_p images from the initial ranking as positive examples. In general, first returned images are highly likely to be relevant, e.g., on Flickr according to [10], [11], in average, among the first 50 returned images at least 37 images are relevant to the query, i.e., 75.37% (estimate obtained for 549 location related queries). Similarly, we consider the last N_n images as negative examples (the very last search results are usually non-relevant). This leads to a total number of N examples ($N = N_p + N_n$) that constitutes an automatic ground truth. The immediate advantage of this strategy is in the complete automation of the relevance feedback process. No real user interaction is actually required, which reduces significantly the processing time as well as the need for conducting complex user studies.

B. *Fisher Kernel visual description*

The entire relevance feedback schemes relies on the efficiency of image content descriptors. We propose a new approach via Fisher Kernel representations. FK [16] represents a signal as the gradient with respect to the probability density function that is a learned generative model of that signal. Recently, [17] introduced the FK as an improved visual vocabulary for Bag-of-Words. Its success shows that it meaningfully captures the visual variation of local descriptors.

We follow [17] and use a Gaussian Mixture Model (GMM) with diagonal covariance matrices as generative distribution. Specifically, let μ_i and σ_i be the mean and standard deviation of the i-th Gaussian centroid, let $\gamma(i)$ be the soft assignment to the i-th Gaussian of the d-dimensional feature x_t captured at image t. The gradient of the GMM with respect to μ_i and σ_i are calculated as [17] (T is the number of images):

$$\mathcal{G}^x_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i} \tag{1}$$

$$\mathcal{G}^x_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]. \tag{2}$$

The final FK representation is achieved by the concatenation of the $\mathcal{G}^x_{\mu,i}$ and $\mathcal{G}^x_{\sigma,i}$ for $i = 1...k$ and has a dimensionality of $2kd$. To employ this framework, images are represented with a set of local descriptors, namely PHOW [15] (dense SIFT descriptors extracted at multiple scales). To make the FK computationally feasible, we apply Principal Component Analysis (PCA) on the original keypoint vectors.

C. *Clustering and pruning*

Equipped with the ground truth and FK content descriptors we use a clustering strategy to group similar appearance images. We selected a Hierarchical Clustering (HC) scheme with a "bottom up" approach (agglomerative)[1]. Besides its low complexity, HC has the advantage of providing a dendrogram of classes by grouping images iteratively based on a certain distance metric. This allows for adapting the number of output classes to the target scenario based on the selection of a cutting point of the dendrogram. HC is applied only to the selected positive and negative examples.

[1] http://www.mathworks.com/help/stats/hierarchical-clustering.html/

TABLE I: Pseudo-relevance diversification results for various descriptors (best results are represented in bold).

descriptor	$P@20$	$CR@20$	$F1@20$
Fisher Kernel (proposed)	0.765	**0.4454**	**0.5558**
CN	0.7374	0.4115	0.5204
LBP	0.7671	0.4188	0.5342
GLRLM	0.7959	0.3943	0.518
CM	**0.8037**	0.4093	0.5343
HoG	0.7549	0.4064	0.5199
CSD	0.7663	0.4216	0.5345
all visual (early fusion)	0.7598	0.4245	0.5349

Once we achieve the clustering, we adopt a supplementary pruning step. A class is declared non-relevant if it contains only negative examples or if the number of negative examples is higher than the positive ones, namely: $N_n^{(i)} \geq 0.5 \cdot N^{(i)}$, where $N_n^{(i)}$ is the number of negative examples in class i and $N^{(i)}$ is the total number of examples in class i. This assumption is based on the fact that cluster images are supposed to be similar with each other. Therefore, if a significant number of negative examples is present, there is a high probability that all the images are in fact negative examples and were assigned wrongly to the positive category.

D. Diversification

The final step is the actual diversification of the results. To enforce the diversity, we restrict the output to contain at least one image from each HC generated cluster. Firstly, for each of the HC output relevant classes (the classes declared as non-relevant are discarded from diversification), the images are sorted according to their initial ranking, so that the first image in a class is the one which has the highest rank in the initial retrieval results. Considering the order described above and starting with the first class, i.e., the class labeled as the first one by the HC scheme, we select as output first ranked image from each class. This leads to N_c images, where N_c is the total number of classes. The process is repeated iteratively, and classes are covered again by selecting the second ranked images, third ranked and so on.

III. Experimental results

Experimental validation was conducted on a publicly available search result diversification benchmarking dataset, Div150Cred [10]. It consists of 153 location related queries (e.g., museums, bridges, parks, monuments, etc) with up to 300 photos per query retrieved from Flickr using Flickr's default "relevance" algorithm (a total of 45,375 images). Images are annotated for both relevance and diversity by human assessors. In particular, for diversity, images are clustered into similar appearance classes. For experimentation, we use 30 queries (8,923 images) for training and the remaining 123 queries (36,452 images) for the actual evaluation.

To assess performance, we use the standard cluster recall at a cutoff at X images ($CR@X$) [12], a measure of how many clusters from the ground truth are represented among the top X results provided by the retrieval system; the precision at X images ($P@X$), a measure of the number of relevant images among the first X ranked results; and their harmonic mean which is the F1-score, $F1@X$. Results are reported as overall average values over all the queries in the dataset. We use a cutoff at $X = 20$ images which simulates the content of a single page of a typical Web image search engine and reflects user behaviour i.e., inspecting the first page of results in priority.

TABLE II: Comparison to relevance feedback approaches (RBF - Radial Basis Function kernel; best results are represented in bold).

RF approach	*feedback*	*descriptor*	$P@20$	$CR@20$	$F1@20$
proposed	pseudo-rel.	FK	0.765	**0.4454**	**0.5558**
Rocchio [4]	relevance	CN	**0.8549**	0.3385	0.4718
Rocchio [4]	diversity	CSD	0.7126	0.3429	0.455
RFE [5]	relevance	CN	0.828	0.3239	0.4526
RFE [5]	diversity	CN	0.787	0.3561	0.4773
SVM-RBF [6]	relevance	GLRLM	0.8508	0.369	0.505
SVM-RBF [6]	diversity	all visual	0.75	0.4086	0.5172
AdaBoost [7]	relevance	GLRLM	0.8077	0.3666	0.4934
AdaBoost [7]	diversity	LBP	0.7463	0.3779	0.4935

To improve more the relevance of the results, we adopted several additional pre-filtering steps. Firstly, we use the Viola-Jones [13] *face detector* to filter out images with persons as main subject. These images are in general non-relevant for the common user. Secondly, we use an *image blur detector* to remove the images which are out of focus. Severely blurred images are in general not satisfactory results for a query. Finally, in particular for this data, we use a *GPS-based filter* which rejects the images that are positioned too far away from the query location, and therefore which cannot be relevant shots for that location.

In what concerns the parameter tuning, preliminary tests were conducted and all parameters were set to optimal values. This configuration is used in the following experiments.

A. Comparison to other visual descriptors

To prove the efficiency of the proposed Fisher Kernel content description scheme, the first experiment consisted on assessing the performance of other visual descriptors, namely: global color naming histogram (CN, 11 values) — maps colors to 11 universal color names: "black", "blue", "brown", "grey", "green", "orange", "pink", "purple", "red", "white", and "yellow"; global Histogram of Oriented Gradients (HoG, 81 values) — represents the HoG feature computed on 3 by 3 image regions; global color moments computed on the HSV Color Space (CM, 9 values) — represent the first three central moments of an image color distribution: mean, standard deviation and skewness; global Locally Binary Patterns computed on gray scale representation of the image (LBP, 16 values); global Color Structure Descriptor (CSD, 64 values) — represents the MPEG-7 Color Structure Descriptor computed on the HMMD color space; and global statistics on gray level Run Length Matrix (GLRLM, 44 dimensions) — represents 11 statistics computed on gray level run-length matrices for 4 directions: Short Run Emphasis, Long Run Emphasis, Gray-Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Gray-Level Run Emphasis, High Gray-Level Run Emphasis, Short Run Low Gray-Level Emphasis, Short Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, Long Run High Gray-Level Emphasis [10].

Results are presented in Table I. The proposed description scheme allows for a boost in diversification, outperforming the best descriptors by more than 2 percentage points (in cluster recall as well as F1-measure).

B. Comparison to relevance feedback approaches

In the following experiment, we compare our results to other relevance feedback approaches from the literature, namely: Rocchio [4] that changes the initial query point according to user's feedback, Relevance Feature Estimation [5] (RFE) that alters the feature representation by assessing features' importance and some classification-based approaches: Support Vector Machines (SVM) [6] and AdaBoost [7], which formulate the relevance feedback as a two class classification of the negative and positive samples. User relevance feedback is simulated with the images' ground truth in a window of 20 images (this is a common setting that allow good results [14]).

We experimented with two situations: (1) feedback is simulated with the relevance ground truth (*relevance*); (2) feedback is simulated with the diversity ground truth by selecting one image from each image class in the initial feedback window (*diversity*). This should allow for more emphasis on the diversification. The approaches were tuned to best performing parameters (best visual descriptor).

Results are presented in Table II. The first observation is the fact that the use of diversified feedback instead of only relevance allows for improvement over the last one. However, regardless the use of actual image ground truth, the best traditional relevance feedback result in terms of $F1@20$ is 0.5172, achieved with SVM and Radial Basis Function (RBF) kernel. This is almost 4 percentage points less than the proposed approach. This percentage is reflected also in the pure improvement of the diversification captured by the cluster recall (CR).

These results are very promising considering the fact that the proposed approach uses automatically generated feedback, while the other relevance feedback approaches use the actual user ground truth information (both for relevance and for diversification).

IV. Conclusions

We addressed the problem of image search result diversification from the perspective of relevance feedback techniques. We proposed a novel perspective that makes the feedback process completely automatic via pseudo-relevance feedback and considers in priority the diversification, instead of the relevance of the results. To cope with the complexity of visual contents, we also introduced a Fisher Kernel representation that allows efficient representations via capturing the variability of keypoint information from the image. The proposed Fisher Kernel pseudo-relevance feedback approach operates on top of an existing retrieval system by improving its results.

Experimental validation on Flickr data shows the potential of this approach. Firstly, the proposed FK description scheme allows to improve the diversification over other state-of-the-art descriptors (e.g., Histograms of Oriented Gradients). Secondly, the proposed pseudo-relevance feedback outperforms other traditional relevance feedback approaches by as much as 4 $F1@20$-score percentage points, even when feedback was diversified and simulated with actual ground truth.

We therefore proved the benefits of the pseudo-relevance assumption in the context of result diversification opening new perspectives for this area of research. Future work will mainly address exploring more complex diversification scenarios, such as the ones involving multi-concept queries where results tends to be less accurate.

References

[1] R. Priyatharshini, S. Chitrakala, "Association Based Image Retrieval: A Survey," in Mobile Communication and Power Engineering, Springer Communications, Computer and Information Science, 2013, pp. 17-26.

[2] S. Rudinac, A. Hanjalic, M.A. Larson, "Generating Visual Summaries of Geographic Areas Using Community-Contributed Images," in IEEE Transactions on Multimedia, 15(4), 2013, pp. 921-932.

[3] R. H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, "Visual Diversification of Image Search Results," in International Conference on World Wide Web, 2009.

[4] J. Rocchio, "Relevance Feedback in Information Retrieval," in The Smart Retrieval System Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs NJ, 1971, pp. 313-323.

[5] Y. Rui, T. Huang, S.-F. Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," in Visual Communication and Image Representation, 10(1), 1999, pp. 39-62.

[6] S. Liang, Z. Sun, "Sketch Retrieval and Relevance Feedback with Biased SVM Classification," in Pattern Recognition Letters, 29, 2008, pp. 1733-1741.

[7] J. Yu, Y. Lu, Y. Xu, N. Sebe, Q. Tian, "Integrating Relvance Feedback in Boosting for Content-based Image Retrieval," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, USA, 2007, pp. 965-968.

[8] G. Cao, J. Y. Nie, J. Gao, S. Robertson, "Selecting Good Expansion Terms for Pseudo-Relevance Feedback," in ACM International Conference on Research and Development in Information Retrieval, 2008.

[9] B. Ionescu, A. Popescu, M. Lupu, A.L. Gînscă, B. Boteanu, H. Müller, "Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset," in ACM Multimedia Systems - MMSys, Portland, Oregon, USA, 2015.

[10] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "Div400: A Social Image Retrieval Result Diversification Dataset," in ACM Multimedia Systems - MMSys, Singapore, 2014.

[11] M.L. Paramita, M. Sanderson, P. Clough, "Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009," in ImageCLEF, 2009.

[12] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," in International Journal of Computer Vision, 57(2), pp. 137-154, 2004.

[13] B. Boteanu, I. Mironică, B. Ionescu, "A Relevance Feedback Perspective to Image Search Result Diversification," in International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, September 4-6, 2014.

[14] A. Bosch, A. Zisserman, X. Munoz, "Image classifcation using random forests and ferns," in IEEE International Conference on Computer Vision (ICCV), ISSN 1550-5499, pp. 1-8, Rio de Janeiro, 14-21 Oct. 2007.

[15] T. Jaakkola, D. Haussler, "Exploiting Generative Models in Discriminative Classifiers," in International Conference on Advances in Neural Information Processing Systems II, ISBN:0-262-11245-0, pp. 487-493, 1998.

[16] F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in European Conference on Computer Vision (ECCV), LNCS 6314, pp. 143-156, 5-11 September, Heraklion, Crete, Greece, 2010.