Towards an Automatic Evaluation of Retrieval Performance with Large Scale Image Collections

Adrian Popescu[1], Eleftherios Spyromitros-Xioufis[2], Symeon Papadopoulos[2], Hervé Le Borgne[1], Ioannis Kompatsiaris[2]
[1]CEA, LIST, 91190 Gif-sur-Yvette, France, {adrian.popescu,herve.le-borgne}@cea.fr
[2]CERTH-ITI, Thermi-Thessaloniki, Greece, {espyromi,papadop,ikom}@iti.gr

ABSTRACT

The public availability of large-scale multimedia collections, such as the Yahoo Flick Creative Commons (YFCC) dataset, facilitates the evaluation of image retrieval systems in realistic conditions. However, due to their size, the creation of exhaustive ground truth would require huge annotation effort, even for limited sets of queries. This paper investigates whether it is possible to estimate retrieval performance in absence of manually created ground truth data. Our hypothesis is that it is possible to leverage existing weak user annotations (tags) to automatically build ground truth data. To test this hypothesis, we implemented a large-scale retrieval pipeline based on two state-of-the-art image descriptors and two compressed versions of each. The top 50 results obtained with each configuration are manually annotated to estimate their performance. Alternately, we produce an automatic performance estimation based on pre-existing user tags. The automatic performance estimations exhibit strong positive correlation with the manual ones and the corresponding system rankings are found to be similar. Hence, we conclude that despite being incomplete and sometimes imprecise, weak user annotations can be leveraged to assess retrieval performance. As a by-product, we release state-of-the-art image features for YFCC and a reusable evaluation package to encourage its use in the community.

Keywords

evaluation, scalability, multimedia collections, YFCC

1. INTRODUCTION

Evaluation has a central role in the development of multimedia retrieval systems since it provides feedback about the quality of the retrieval results of such systems and typically offers valuable insights about ways to improve them. A large number of evaluation methodologies were proposed and tested principally via evaluation campaigns such as ImageCLEF [13], Pascal [4], MediaEval [13], and ILSVRC [17]. An important objective of such initiatives is to propose robust evaluation testbeds that can be used to evaluate systems in a reproducible fashion over publicly available datasets. In most cases, the evaluation is based on the manual construction of *reference golden standards* (typically referred to as *ground truth*) that is associated to datasets. While feasible at small and medium scale, thorough multimedia annotation is necessarily partial for larger scale datasets. For instance, image classification datasets such as PascalVOC 2007 [4] and ImageCLEF Photo Annotation 2012 [21] have a complete annotation of 20 and 95 concepts for approximately $10,000$ and $25,000$ images respectively. At larger scales, ILSVRC uses the pre-existing ImageNet annotations that provide only one label per image for a classification evaluation dataset that includes over 1.2 million images. In image retrieval, the ImageCLEF Wikipedia Retrieval task [23] implemented a pooling approach for a collection of over $237,000$ images. The top results obtained for each topic by each participating system were merged and a partial ground truth was obtained for the image collection.

There are fewer cases when the evaluation methodology is specifically designed to exploit pre-existing information. For instance, the MediaEval Placing Task [2], of which the objective is to evaluate the accuracy of automatic geotagging, reuses GPS information associated to images as ground truth. Another such example is the KBP Entity Linking task [11], which exploits Wikipedia links as ground truth, and requires participants to associate the right concept with an entity mention. Towards this direction, we study the feasibility of repurposing weak user annotations (tags) for the evaluation of image retrieval systems. Previous studies [24] showed that user annotations are incomplete and often imprecise. We illustrate these two problems in Figure 1. Assuming that a user's information needs are expressed by the queries listed above the images, the two examples to the left of the figure illustrate incompleteness. While relevant for the content, `car` and `feline` are not among the tags of the images. However, these two images would be accurate results if one performed a textual query using the terms `Mini Cooper` and `black cat` respectively (assuming that the search engine is looking for those terms on the set of tags). The two examples to the right of the image illustrate imprecise tagging if one assumes that the user's information needs are *boxing* as *sport* and *Christmas tree*.

In spite of tagging problems, we will show empirically that user tags can be successfully used to estimate the retrieval quality of different systems. While the derived performance measures may be inaccurate, systems are ranked in the same

MMCommons'15, October 30 2015, Brisbane, Australia
©2015 ACM. ISBN 978-1-4503-3744-1/15/10...$15.00
DOI: http://dx.doi.org/10.1145/2814815.2814819.

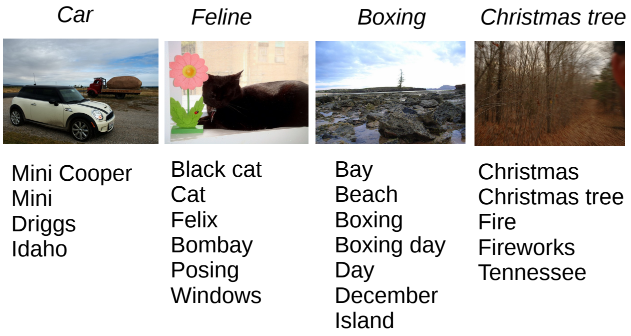

Figure 1: Illustration of tagging problems. Information needs (queries) are presented above the images and user annotations below them. The two images to the left illustrate tagging incompleteness and the two images to the right illustrate imprecision.

order that would come out of a result pooling evaluation approach. In addition to evaluation, our paper also presents a thorough comparative study of state-of-the-art features. We notably compare pre-CNN (*VLAD+SURF*) and CNN (*VGG*) features and show that the latter obtain better performance on a very large collection. We also present results with PCA-compressed versions of each feature (to 128 and 16 dimensions) and find that, while more compact, the 128 dimensional versions have comparable performance to the one obtained using the full versions of the features.

2. RELATED WORK

2.1 Image retrieval evaluation

Publicly available datasets are necessary for result reproducibility in image retrieval. The effort needed to create comprehensive ground truth annotations for small datasets is manageable but increases linearly with dataset size. Scale is an important factor in evaluation since larger datasets are more representative of real-life retrieval, as in the case of Web image search engines. As mentioned above, ground truth annotations are often created manually and they can be either complete or incomplete. For complete ground truth annotations, all collection images are annotated against a full set of query concepts. In this category, PascalVOC 2007 [4], MIR Flickr [6] and NUS-WIDE [3] are annotated with 20, 24[1] and 81 concepts, and include $9,963$, $25,000$ and $269,648$ images respectively. At a larger scale, ImageNet [17] includes single-concept annotations for $21,841$ concepts that are illustrated with a total of $14,197,122$ images. One of the limitations of ImageNet is that they include annotations for single concepts, while image retrieval queries often contain a combination of concepts [23].

Datasets dedicated to image retrieval were created in specially designed campaigns, including the recurring ImageCLEF Photo Retrieval [5] and Wikipedia Retrieval tasks [23]. The IAPR TC-12 dataset [5] includes $20,000$ diversified tourism images. Annotations are of high quality and thus do not reflect very well the noisy character of Web collections. The latest versions of the Wikipedia collection associated to this task, dating from 2010 and 2011, include partial annotations of 70 and 50 topics that include one or more concepts. The ground truth is obtained by pooling the top results of participating systems, implying that only a part of the collection is judged against each query. While more scalable than full annotation, this approach does not allow a fair evaluation of new methods on the same dataset since all results that were not seen during pooling would be always considered as irrelevant.

An alternative path towards scaling up retrieval benchmark datasets is to combine a fully annotated small scale dataset and a large unlabeled background collection [10]. In such cases, it is assumed that all images from the background collection are irrelevant (also referred to as *distractors*) and the performance is measured by how well each method preserves the order of relevant images from the initial collection. This approach has been often tried with selected datasets, e.g., the INRIA Holidays collection [8], for evaluating instance/near-duplicate retrieval systems under the assumption that most instances will appear very rarely or not at all in the background collection. However, it is less suited for ad-hoc retrieval because it penalizes topics that appear frequently in the background collection. For instance, if one considers a topic such as `black cat` present in a small scale collection, it is likely that many more relevant results for this topic are found in the background collection.

2.2 Descriptors for image retrieval

Proposing a new descriptor for image retrieval is beyond the scope of this paper. Instead, we focus on the retrieval performance that is achieved by systems using pre-CNN and CNN features.

Until recently, state-of-the-art image retrieval methods relied on local feature aggregation approaches. In such approaches, hand-crafted descriptors such as SIFT [14] and SURF [1] are first extracted from local image patches. These descriptors are then aggregated into a single vector that summarizes their statistics and is used as the global image descriptor. A notable example, is the popular Bag-of-Words (*BoW*) representation [19] where local descriptors are assigned to words of a visual vocabulary and the final vector is the histogram of the distribution of visual words in the image. In an attempt to improve the quality of *BoW*, a number of new representations emerged that encode higher-order statistics of the distribution of features to visual words. Characteristic examples are the *VLAD* [9] and the *Fisher* vector [15]. Both *VLAD* and *Fisher* vectors have shown significantly better performance than *BoW* in image retrieval tasks and were considered as state-of-the-art until 2012 [10].

Many computer vision tasks, including object classification and localization, are currently tackled with Convolutional Neural Network (*CNN*) architectures. The ImageNet 2012-2014 challenge methods and results [17] are illustrative for this shift towards the use of *CNN*. While their use for image classification is dominant, *CNN*-based descriptors have been also exploited for image retrieval. For instance, a generic *CNN* model was shown in [16] to provide good performance on different small-scale evaluation datasets. Here, we exploit the *VGG* features [18], a representative of *CNN* descriptors, that achieved the second best classification performance at ILVRSC 2014.

[1] In addition to the 24 concept annotations originally provided with the MIR Flickr dataset, the ImageCLEF Photo Annotation 2012 campaign [21] made available annotations on 95 concepts.

3. EVALUATION METHODOLOGY

We propose an evaluation methodology that follows the same protocol as existing ones, except for the creation of ground truth annotations, which, here, are substituted by pre-existing weak user annotations. Our hypothesis is that, although imperfect, pre-existing user annotations can be reliably used to compare content-based image retrieval (CBIR) systems. As we mentioned, the two main annotation imperfections that need to be handled are *imprecision* and *incompleteness*. Imprecision, i.e. tags that are wrongly associated to a given image, is easy to deal with, since images that are wrongly annotated with a topic are unlikely to appear among the top results associated with a topic. Assume that we use one or several example images of *boxing* (the sport) to retrieve images of this topic in a large collection. The visual neighbors found for this topic are unlikely to be images such as the one tagged with `boxing` in Figure 1. Ground truth incompleteness refers to the fact that not all images that are representative for a concept are actually tagged with it. In Figure 1, the image that represents a *car* is not tagged with this concept although it is representative of it. Incompleteness is thus a more serious problem than imprecision because a part of the relevant images that are retrieved will not be considered as such. However, our hypothesis is that incompleteness will equally affect all compared methods and, if the evaluation is done appropriately, the ranking of these methods will be similar to the one that would be obtained with manually created ground truth annotations.

The evaluation is designed following a standard evaluation protocol for CBIR [23]. The process comprises the following steps: topic selection, ground truth creation, and performance analysis. The validation of our automatic CBIR evaluation is done by creating ground truth annotations automatically and manually in order to be able to compare the results obtained in these two ways.

3.1 YFCC Collection

YFCC [22] is currently the largest publicly available multimedia collection. It includes a total of 99.2 million images and 0.8 million videos from Flickr. All contents are licensed under different versions of Creative Commons and were uploaded between 2004 and 2014. Due to its size and diversity YFCC covers a wide spectrum of topics and enables research that focuses on both performance and scalability of tested multimedia retrieval methods. Here, we focus on the still images contained in YFCC in order to support CBIR over the collection. The images are not publicly available and hence we had to download them in April 2015. Out of the full dataset, 96.7 million were still available at that time and were used in the experiments of Section 4.

3.2 Retrieval Pipeline

We implement a classic CBIR pipeline. The images are represented using *VLAD+SURF* [20] and *VGG* [18], two basic features that are characteristic of pre-CNN and CNN approaches respectively. Compressed versions of these features are used to create supplementary retrieval configurations. Given a query image, its nearest neighbors are retrieved by computing its L^2 distance to the collection images. The retrieval process is performed per topic and, if there is more than one query image, the nearest neighbors of the topic are obtained by combining the top lists of nearest neighbors for individual image queries. The most similar images are naturally favored and if one collection image appears for more than one example image, it is kept only with the lowest L_2 distance (i.e. its highest rank).

3.3 Image Descriptors

We extracted the following state-of-the-art image features for the YFCC collection:

- *VGG* was proposed by [18] for the the ILSVRC 2014 challenge and the respective system was ranked second best in the classification task. The features are extracted with a 16-layer CNN and we exploit the output of the last fully connected layer (*fc7*), which consists of $4,096$ dimensions. This layer is selected since its neurons convey a relatively high-level encoding of the image content that is well suited for finding semantically similar images for a query.

- VGG_{PCA}. Direct CBIR with full *VGG* features has significant computational cost in large scale collections, such as YFCC, since a comparison of 4096-dimensional features is needed. To speed-up the retrieval process, we create a PCA compressed version of *VGG*. The PCA matrix is computed using a sample of $250,000$ YFCC images, randomly selected from the collection. We report results obtained with the most significant 16 and 128 dimensions from the PCA representation of *VGG*, denoted respectively as VGG_{16} and VGG_{128}.

- *VLAD*. Improved *VLAD* vectors [20] using *SURF* and the multiple vocabulary aggregation technique [7] with four visual vocabularies (of $k = 128$ centroids) were extracted. Their initial dimensions $(32,768)$ were reduced with PCA+whitening. $VLAD_{1024}$, $VLAD_{128}$ and $VLAD_{16}$ correspond to retaining the 1024, 128 and 16 most significant dimensions respectively.

3.4 Topic Selection

The appropriate selection of topics is a core requirement for a successful evaluation [23]. The topic set needs to meet all the following constraints:

- Representativeness - include topics that are illustrated in the test collection. Even though the test collection is very large, it would be useless to perform retrieval for topics that are not represented in the collection.

- Diversity - cover different conceptual domains and to be neutral with respect to the tested method. For instance, if one of the methods is very robust to geometrically invariant images, this type of topics should represent only a fraction of the topic set.

- Robustness - include enough topics to ensure stable ranking of the tested methods. Naturally, stabler results are obtained with larger topic sets but the manual annotation effort required for the ground truth creation increases linearly with the number of topics.

- Difficulty - topics should cover a wide range of difficulty levels. While it is not straightforward to predict difficulty in advance, this aspect can be estimated by analyzing the visual complexity of the topic and the number of images tagged with the topic.

To meet the above constraints, we select a set of topics that are present in the test collection and range from simple to difficult. The topics are further described in the next subsection.

3.5 Ground Truth Creation

3.5.1 Automatic Ground Truth

We selected 50 topics that were sampled from a total of 120 topics that are part of the ImageCLEF Wikipedia 2010 and 2011 evaluation tasks [23]. This choice is made because Wikipedia Retrieval topics are diversified and are also checked against search engine logs, hence they are illustrative of real Web queries. The average YFCC topic frequency is $47,405$, with a standard deviation of $76,914$. The topics that are most frequent in YFCC are: *graffiti* ($314,163$ occurrences); *airplane* ($281,973$); bridge ($280,848$). The rarest topics from our test set are: *tennis player* (510); *Chinese characters* (292) and *music sheet* (108).

Since incompleteness was identified as a problem for the constitution of the automatic ground truth, effort was put to reduce it by adding synonyms, as well as plural forms of the topics. Synonyms were extracted from Wikipedia and WordNet and only those whose main sense is used to describe the topic were retained. For instance, images associated with *airplane* were obtained by matching YFCC tags against the query: `airplane` OR `airplanes` OR `aeroplane` OR `aeroplanes` OR `plane` OR `planes`.

3.5.2 Manual Ground Truth

The manual ground truth is obtained using a pooling approach. For each topic and each tested method, the top 50 results are selected and manually assessed by an expert annotator. The pooling depth of 50 was chosen, because it was shown to be sufficient for obtaining stable rankings in [23]. The annotator first read the topic name, an associated narrative that describes the information need and saw three image examples that illustrate the expected content. Then, the annotator went through all images associated with the topic and selected only those that were judged to be representative of it. Note that images retrieved by competing methods were presented randomly in the annotation interface. Aggregating the images returned by the six competing systems (corresponding to the different tested features) and assessed by the annotator, the resulting average number of relevant images per topic was found to be 145.7, with a standard deviation of 96.3.

4. PERFORMANCE ANALYSIS

The analysis of results is carried out using average precision scores at different recall levels. These recall levels are useful in order to assess the stability of rankings. Table 1 illustrates the results obtained using the different descriptors described in Section 3.3 with the manually and automatically constituted ground truth annotations. The rankings obtained with the two types of ground truth annotations are identical for all recall levels up to P@500. Naturally, due to the incompleteness of the automatic ground truth, the associated scores are roughly four times lower than the manual precision for P@50.

To measure the association between the automatically estimated and the manual performance scores, we calculated the Kendall rank correlation coefficient [12] and performed the corresponding non-parametric hypothesis test for statistical dependence. This was done both at topic level, by pairing manual and automatic scores for all distinct topic-system combinations (300 pairs), and at system level by pairing the average (across all topics) manual and automatic precision scores of each system. According to these tests (summarized in Table 2), all variants of the automatic metric are statistically dependent with the manual metric at both the topic and the system level.

Table 2: Kendall rank correlation coefficient and corresponding p-values between manual and automatic scores.

Prec. level	Topic level	System level
@50	0.547 $p < 0.001$	1.000 $p = 0.003$
@100	0.572 $p < 0.001$	1.000 $p = 0.003$
@200	0.583 $p < 0.001$	1.000 $p = 0.003$
@500	0.585 $p < 0.001$	1.000 $p = 0.003$
@1000	0.587 $p < 0.001$	0.867 $p = 0.017$
@2000	0.578 $p < 0.001$	0.867 $p = 0.017$

We also conducted an analysis of statistical significance between the P@50 results obtained with the manual and automatic ground truth annotations and present the results in Figure 2. This analysis is carried out at topic level using a two tailed t-test. It is performed separately for all pairs of results obtained with manual and automatic ground truth annotations respectively. In an ideal case, the two matrices would be identical and, while this is not the case, the corresponding p-levels are identical in the majority of cases. For instance, the results obtained with VGG are significantly different ($p < 0.1$) from all other features except for VGG_{128} with both ground truth annotations. We note that the use of manual ground truth leads to better separability between the tested methods. The results illustrated in Figure 2 indicate that a larger number of topics should probably be used with automatic ground truth annotations in order to improve the separability of the methods.

Beyond P@500, the results start to diverge, notably for the comparison of $VLAD_{1024}$ and $VLAD_{128}$. If the automatic evaluation procedure is used, recall levels between 50 and 500 should be taken into consideration since they offer the best correlation between the manual and automatic scores. Our hypothesis concerning the use of automatically constituted ground truth annotations for estimating retrieval performance is confirmed. This finding is important since it facilitates the use of very large scale and realistic image datasets without the burden of manual evaluation.

To our knowledge, this is the first time CNN features are evaluated for retrieval over a collection as large as YFCC. The results presented in Table 1 reveal that they are highly effective for image retrieval. The performance attained using VGG and VGG_{128} is approximately three times higher than the one using $VLAD_{1024}$ and $VLAD_{128}$ at low recall levels (P@50, P@100). It is also noticeable that VGG_{16}, a highly compressed version of VGG also provides good results, roughly twice as good as the ones of $VLAD_{128}$. This performance difference is likely due to the fact that $VLAD$ and other local feature based approaches are mainly tailored for instance and near-duplicate image retrieval. As a result, their inferior performance on topic-based retrieval is to be expected. VLAD still performs better than VGG on queries such as *Sagrada Familia* where the topic is narrow (in terms of visual appearance). An equally interesting observa-

Table 1: Results obtained with the tested descriptors. The manually computed precision score (P@50 man) is given on the second column, followed by the automatically computed scores at different recall levels.

	P@50 man	**P@50**	**P@100**	**P@200**	**P@500**	**P@1000**	**P@2000**
VGG	**0.626**	0.1892	0.1682	0.1568	0.1405	0.1308	0.1213
VGG_{128}	**0.610**	0.1756	0.1638	0.1500	0.1368	0.1257	0.1168
VGG_{16}	**0.438**	0.1180	0.1002	0.0933	0.0837	0.0771	0.0706
$VLAD_{1024}$	**0.192**	0.0656	0.0508	0.0404	0.0298	0.0242	0.0202
$VLAD_{128}$	**0.217**	0.0664	0.0520	0.0415	0.0305	0.0241	0.0197
$VLAD_{16}$	**0.084**	0.0260	0.0184	0.0138	0.0107	0.0089	0.0081

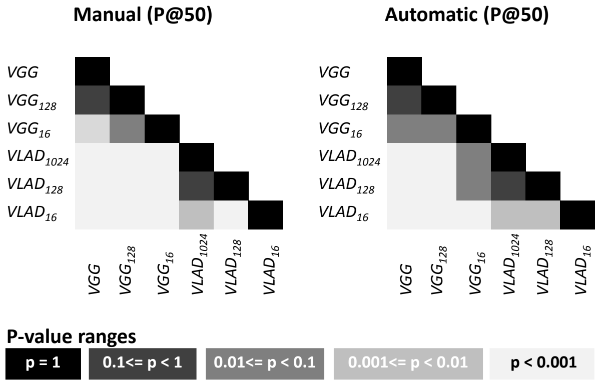

Figure 2: Analysis of the statistical significance of P@50 results for the manual (left) and automatic (right) ground truth annotations. The results are grouped in p-value intervals. The lighter the gray level, the clearer the statistical independence between the two compared sets of results.

tion is that the relative performance drop of VGG features when increasing recall is smaller than that of $VLAD$. For instance, the precision achieved with VGG drops by approximately 50% between P@50 and P@2000 whereas the one obtained with $VLAD_{128}$ drops by more than 200%. While simple, feature compression with PCA indicates that it is highly effective for both VGG and $VLAD$. For VGG, the performance loss between 4096 dimensions and 128 reaches 2.2% with the manual ground truth and is estimated between 2.9% (P@500) and 6.4% (P@50) with the automatic ground truth. Somewhat surprisingly, for $VLAD$, the 128 dimensions version of the feature has slightly better performance than the one obtained with the 1024 version.

Figure 3 illustrates retrieval results for six topics by presenting the top five images obtained with the best performing pre-CNN and CNN features, $VLAD_{128}$ and VGG respectively. To illustrate the merits of each method, we selected the top three performing topics for each of them. $VLAD_{128}$ produces high quality results for *Sagrada Familia*, *musician* and *airplane*, with P@50 scores of 0.92, 0.84 and 0.82 respectively. We note that VGG also exhibits high precision for *musician* and *airplane*. A lower score is obtained on *Sagrada Familia*, where the top results obtained with VGG are dominated by panoramic views of cities that do not necessarily include *Sagrada Familia*. Other topics with low scores for VGG include: *Coca Cola* (P@50 = 0.08); *playing cards* (P@50 = 0.08); *chinese characters* (P@50 = 0.02) and *carnival* (P@50 = 0.02). A common characteristic of these topics is that they are not well represented by the 1000 ImageNet concepts used to train VGG. It would likely be possible to further improve results obtained with CNN if the model was learned with a higher number of classes and if these classes would better match the content of the test collection. VGG returns high-quality results (P@50 = 1) for 10 topics, including *polar bear*, *baseball* and *butterfly* that are illustrated in Figure 3. $VLAD$ has lower scores for these topics, especially for *polar bear*, with P@50 = 0.

5. CONCLUSION AND FUTURE WORK

We presented an evaluation methodology that justifies the usage of large tagged image collections for retrieval system comparison based on the pre-existing user tags. The evaluation shows that method rankings obtained with automatically and manually obtained ground truth annotations are identical, thus validating our scalable and cost-effective evaluation approach. The results reported here encourage us to pursue work in the following directions: (1) make the manual ground truth more robust by increasing the number of topics and by performing validation with several annotators; (2) confirming the obtained results with other evaluation measures aside precision; (3) test supplementary visual features; (4) artificially inject supplementary noise to test the robustness of the evaluation; (5) test the evaluation methodology on other datasets.

While the proposed evaluation methodology is general for CBIR, it also has some important limitations. First, it would be difficult to apply it to text or text-image retrieval due to the fact that it considers all images tagged with a topic name as correct. Second, while the significance analysis shows that the correlation between manual and automatic scores is generally high at system level, it is far from perfect.

In addition, we have also performed a thorough comparison of pre-CNN and CNN features over a very large multimedia collection. Our results confirm those obtained on smaller databases in that CNN features have clearly superior performance. Equally important, we empirically found that both types of features can be compressed, thus leading to a more efficient representation, with little or no performance loss. While the tested features do not allow one to perform real-time queries on the entire dataset, scalability could be further improved by applying hashing approaches such as Product Quantization.

In order to stimulate future research on YFCC, we release an evaluation package, as well as the 128-dimensions versions of the VGG and $VLAD$ features, along with the PCA matrices needed to produce them[2].

[2] http://mklab.iti.gr/project/vlad-vgg-yfcc-dataset

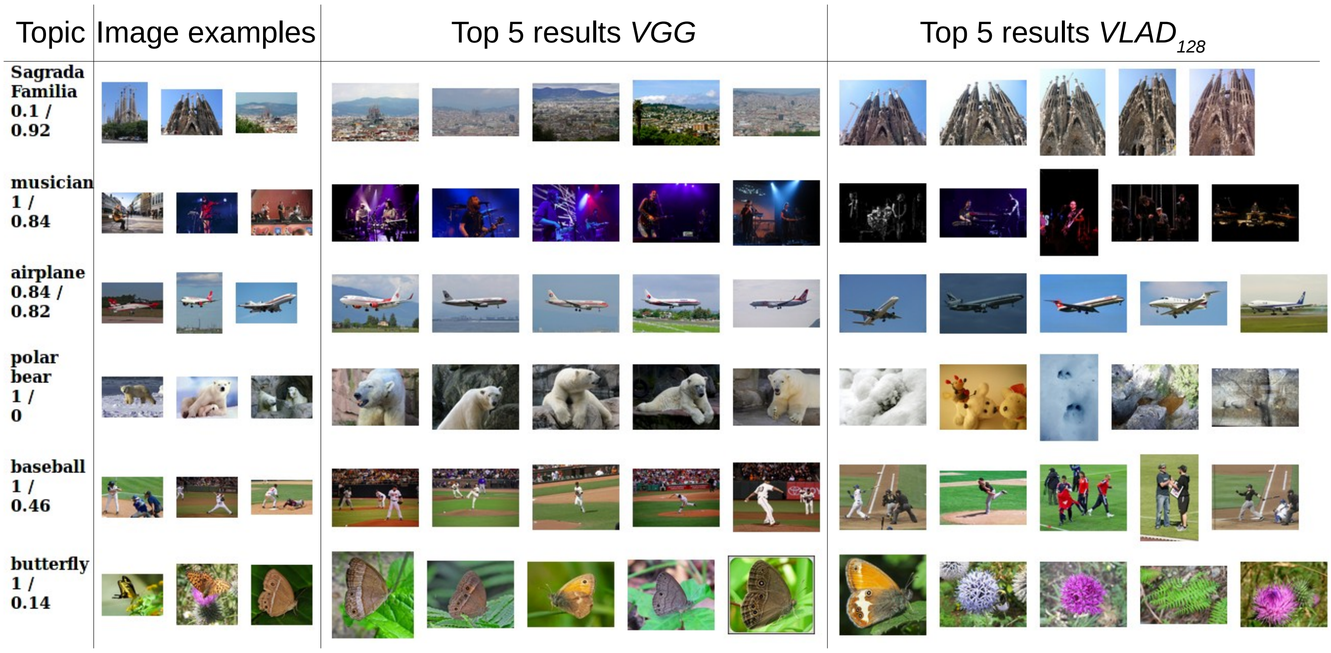

Figure 3: Illustration of retrieval results for six topics, those that have the highest manual P@50 performance for $VLAD_{128}$ and VGG. Below the topic name, we also present the P@50 scores for VGG and $VLAD_{128}$. The three image examples are presented for each query. The first three rows illustrate the topics with best scores for $VLAD_{128}$ and the others, three of the best ranked topics for VGG.

6. ACKNOWLEDGMENTS

This work is supported by the USEMP FP7 project, partially funded by the EC under contract number 611596.

7. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[2] J. Choi and al. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *GeoMM 2014 workshop*.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *ACM CIVR 2009*.

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015.

[5] M. Grubinger, P. Clough, H. MÃijller, and T. Deselaers. The iapr benchmark: A new evaluation resource for visual information systems. In *LREC*, Genoa, Italy, May 2006.

[6] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In *Proc. of ACM MIR 2010*.

[7] H. Jégou and O. Chum. Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In *Computer Vision–ECCV 2012*, pages 774–787. Springer, 2012.

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

[9] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR 2010*. IEEE Computer Society.

[10] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE PAMI*, 34(9):1704–1716, 2012.

[11] H. Ji, R. Grishman, and H. Dang. Overview of the TAC2011 knowledge base population track. In *TAC 2011 Proceedings Papers*, 2011.

[12] M. G. Kendall. Rank correlation methods. 1948.

[13] M. A. Larson and al, editors. *Working Notes Proceedings of the MediaEval 2014 Workshop*, volume 1263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[15] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

[17] O. Russakovsky and al. Imagenet large scale visual recognition challenge, 2014.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

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

[20] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. A comprehensive study over vlad and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 2014.

[21] B. Thomee and A. Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. In *CLEF 2012 Evaluation Labs and Workshop, Working Notes*, 2012.

[22] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[23] T. Tsikrika, J. Kludas, and A. Popescu. Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. *IEEE MultiMedia*, 19(3):24–33, 2012.

[24] A. Znaidia. Handling imperfections for multimodal image annotation. *PhD thesis*, 2014.