Exploring the Impact of Inter-query Variability on the Performance of Retrieval Systems

Francesco Brughi[1(✉)], Debora Gil[1], Llorenç Badiella[2], Eva Jove Casabella[3], and Oriol Ramos Terrades[1]

[1] Department Ciències de la Computació, Computer Vision Center, Univ. Autónoma de Barcelona, Barcelona, Spain
fbrughi@cvc.uab.es
[2] Servei de Estadística Aplicada, Univ. Autónoma de Barcelona, Barcelona, Spain
[3] Department História i História de l'Art, Univ. de Girona, Girona, Spain

Abstract. This paper introduces a framework for evaluating the performance of information retrieval systems. Current evaluation metrics provide an average score that does not consider performance variability across the query set. In this manner, conclusions lack of any statistical significance, yielding poor inference to cases outside the query set and possibly unfair comparisons. We propose to apply statistical methods in order to obtain a more informative measure for problems in which different query classes can be identified. In this context, we assess the performance variability on two levels: overall variability across the whole query set and specific query class-related variability. To this end, we estimate confidence bands for precision-recall curves, and we apply ANOVA in order to assess the significance of the performance across different query classes.

1 Introduction

An effective performance measure is of essential importance in the development of new learning algorithms. In the case of content-based image retrieval (CBIR), the standard evaluation protocol consists of defining an image query set, computing a performance score for each single query, and finally aggregating - usually averaging - them to obtain a global score. Whereas this is a very compact way to represent and compare algorithm performances, it might not be fully informative since the single global score does not take into account performance variability. In order to estimate if there are significant differences in evaluation scores, a usual practice is to compute confidence intervals for the achieved score. In the context of classification problems, the usage of *bootstrapping* has been advocated [1]. The application of this technique to precision-recall (PR) curves and receiver operating characteristic (ROC) curve is discussed in [2] and [3], respectively. Bootstrapping basically consists in repeatedly taking random samples, with replacement, from the data points (images from the test sets, in our case).

A. Campilho and M. Kamel (Eds.): ICIAR 2014, Part I, LNCS 8814, pp. 413–420, 2014.
DOI: 10.1007/978-3-319-11758-4_45

Fig. 1. Examples of four different motive classes

From each sample, a curve will be generated. Alternatively, cross-validation can be used to repeatedly split the dataset into a training and a test set. This produces a curve for each split. Once multiple curves are obtained from the data, several methods exist in order to generate confidence bands for each curve [4]. The variability caught by this approach is entirely associated to the search space, as it depends on the test dataset images. In the context of CBIR, aside from the variability associated to the whole search space [2], there are specific variability factors associated to each query. As a matter of fact, this variability is lost when averaging the individual query scores in order to obtain an overall measure (such as mean average precision).

Variability is particularly critical in the case of a very heterogeneous set of queries, given that the algorithm performances is prone to vary significantly across the query set. This is the case, for instance, of artistic motive retrieval from ancient Greek pottery digital repositories [5], [6]. In this context, we have a set of queries divided into several classes (some examples in Figure 1) which, as discussed in [5], show high inter-class variability. The exploratory study presented in [5] also showed that the method that best performs on a certain query class, might not be as effective on the others. In this context, evaluating a system with an overall score which aggregates the individual query results does not provide enough information to select the best solution. Since the interest is to assess the robustness of the tested methods, this motivates to produce an evaluation metrics capable to capture the method average performance as well as its variability when applied to different query classes. In this direction, besides the context of image processing, a large amount of work has been published in the field of test diagnostics concerning the estimation of test scores - such as the area under the receiver operating characteristic curve (AUC) - and their variability when comparing different scores. Both non-parametric [7], [8] and parametric approach, based on normality assumption [9], have been proposed in the literature. A common concern is the impossibility of these methods to analyse the sources of variability and the factors influencing the performance of a system.

This paper presents a statistical framework that allows us to evaluate and compare different CBIR methods, in terms of the factors that most influence their performances. Our evaluation scheme is focused on studying the performance variability associated to the different classes of query as well as allowing for a class-wise comparison. Our comparison framework has been applied to 2

standard methods and experiments show the influence of the query type in their performances.

2 Assessment of Inter-query Variability

The common evaluation protocol for CBIR, inherited from information retrieval [10], is based on the notions of *relevant* and *non-relevant* retrieved images for a certain query. Given a test set and a set of query images, for each query a CBIR system is asked to output a number of ranked list of the test set images, according to a relevance measure of the images to the query. The quality of the ranked lists is evaluated based on whether the first k retrieved images are actually relevant or not for the given query. Whereas in binary classification problems *true positive rate* (TPr) and *false positive rate* (FPr) are commonly used, the standard evaluation metrics in CBIR are *precision* and *recall* (also known as *sensitivity*) since they better deal with unbalanced class distributions, which are typical in retrieval tasks [11]. Precision $p(k)$ and recall $r(k)$ for the first k elements of the output ranked list are defined as

$$p(k) = R(k)/k \qquad \text{and} \qquad r(k) = R(k)/N_{rel}, \tag{1}$$

where $R(k)$ is the number of relevant documents contained in the top k ranked elements and N_{rel} is the total number of relevant documents contained in the test set. Precision measures how many of the retrieved documents are actually relevant for the query, whereas recall estimates how many of the relevant documents have been retrieved. The plot given by precision and recall values obtained for each query, called *precision-recall* (PR) *curve*, is commonly used to visually assess the CBIR systems. For each query, the area under the PR curve, known as average precision (AP), is the usual evaluation score of the single query retrieval, and it is given by $AP = \frac{1}{2}\sum_{k=2}^{N}[p(k) + p(k-1)][r(k) - r(k-1)]$. The overall system performance score is then computed by averaging the AP values obtained for each query. This score is known as *mean average precision* (mAP), and it is normally used to compare the performances of different algorithms, given a query set and a test set.

As pointed out in Section 1, mAP comparisons might yield unfair results and cannot detect the sources of error and variability in performance. The PR curves and the corresponding APs will be used in the following for our study on CBIR system evaluation. As introduced in Section 1, we are interested in estimating the performance variability within query sets (or subsets such as classes) in order to achieve a more informative evaluation of a retrieval system. Quantifying the variability of the performance for different queries within a set can be useful to assess the method robustness for that set. Such variability can be obtained by exploring the differences on PR curves and APs across a given set. Variability of PR curves will be assessed by computing confidence bands for curves sampled over a given population group. Confidence bands will be computed using *vertical averaging* (VA) [4]. VA consists of stacking precision values from the different samples that correspond to the same recall values. Therefore, the precision has to

be expressed directly as a function of the recall. This can be done by obtaining k from (1) as $k(r) = R^{-1}(rN_{rel})$. It must be noted that R is monotonically increasing within its domain, which guarantees the existence of its inverse R^{-1}. By substitution, we find $p(r) = rN_{rel}/R^{-1}(rN_{rel})$. In practice, $p(r)$ is only defined for a discrete set of recall values within $[0, 1]$, which vary across different queries. Therefore, we linearly interpolated the function in $[0, 1]$ and we sampled the recall with step $1/(N_P - 1)$, where N_P is the number of quantiles. For each sampled quantile, the average defining the confidence band is computed from a given a set of N_Q query images, thus, N_Q PR curves, as follows. Let $p_j^q = p^q(r_j)$ be our precision observations for the j-th quantile, $j = 1, \ldots, N_P$, and the q-th query image, $q = 1, \ldots, N_Q$. If μ_{p_j}, σ_{p_j} are, respectively, the unbiased sample "vertical" mean and variance for the j-th quantile, then the interval for μ_{p_j} at confidence level $1 - \alpha$ is:

$$\left[\mu_{p_j} - t_{\alpha/2}^{N_Q-1}\frac{\sigma_{p_j}}{\sqrt{N_Q}}, \quad \mu_{p_j} + t_{\alpha/2}^{N_Q-1}\frac{\sigma_{p_j}}{\sqrt{N_Q}}\right], \tag{2}$$

where $t_{\alpha/2}^{N_Q-1}$ is the value of a t-Student distribution with $N_Q - 1$ degrees of freedom. Joining the confidence intervals computed for all the N_P quantiles, we obtain the confidence band of the overall curve.

Confidence bands already provide visual assessment for significance difference in performance for 2 CBIR systems. In order to numerically check whether a method performance significantly differs across query classes, we will use *analysis of variance* (ANOVA) [12]. ANOVA is a statistical tool used to test data when it consists of a quantitative response variable and one or more categorical explanatory variables (or factors). In its simplest form, it allows to check the hypothesis that all the groups (corresponding to the different factors) have the same population mean. In our case, we want to study the different performances between different query classes as well as between different methods. Therefore our factors will be all possible method-query class pairs, whereas an intuitive choice for the response variable is constituted by the AP. We will denote by N_C the number of query classes, and by n_c the number of images belonging to the c-th query class, being $c = 1, \ldots, N_C$. Assuming that we want to compare 2 methods, A and B, our factors are defined as $X^{c,m}$, where m is either A or B. The response variable, i. e. the AP score for the q-th query and the method m, will be represented by $Y^{c,m}$. This way, for each ANOVA group - defined by the factor $X^{c,m}$ and the response variable $Y^{c,m}$ - we have n_c observations $\{\hat{Y}_q^{c,m} : q \in \mathcal{C}_c\}$, being $\mathcal{C}_c$ the set of all subscripts q that belong to the c-th class. Then, we can express the ANOVA null hypothesis as

$$H_0 : \mu_{Y^{1,A}} = \ldots = \mu_{Y^{N_C,A}} = \mu_{Y^{1,B}} = \ldots = \mu_{Y^{N_C,B}}, \tag{3}$$

which states that the precision observations obtained for the N_C query classes and the 2 different methods come from distributions with the same mean.

The ANOVA outcome indicates whether it is possible to reject the null hypothesis or not. Yet, what we are interested to know is, for instance, which is the best performing method-class combination, or whether there is a significant

difference between two specific performances. We can answer these questions by applying pairwise comparison to the ANOVA outcome. In particular, we have used Tukey's *honestly significant difference test* (HSD) [13], which compares the difference between each pair of factors with appropriate adjusting for multiple testing. HSD is similar to a t-test, except that it takes into account the fact that when there are multiple comparisons being made, the probability of making a type I error increases [13]. Given a pair of factors, after estimating their $1 - \alpha$ confidence intervals, the test considers them significantly different if their intervals are disjoint, and not significantly different otherwise.

3 Experimental Set-Up

The goal of these experiments is to assess the impact of variability in performance evaluation of retrieval systems using the methods described in Section 2. We have chosen the well known Oxford 5k dataset[1], which contains 5062 images of building "landmarks" from different viewpoints. A landmark is intended to be a particular of a building. The landmarks are divided into 11 classes. Ground truth is provided as follows. For each class, 5 images are annotated as queries. The remaining images are annotated as: *good* if the landmark is clearly visible, *ok* if more than the 25% of the landmark is clearly visible or *junk* if less than the 25% of the landmark is visible or distortions are present, *absent* when the landmark does not appear. Given that the number of images for the different classes is highly variable (considering together *good* and *ok*, it ranges from 7 to 220), we selected a subset with balanced number of elements per class, since we do not want the dataset imbalance to affect our statistical analysis. Our subset of the Oxford 5k was created as follows. We picked the 5 classes that have the highest numbers of elements (Fig. 2), among the *good* and *ok* annotated images. Using the minimum of these numbers, we randomly sampled each class, without replacement, until obtaining 5 subclasses with the same number of images. Then, we added 300 distractor images, randomly sampled among the ones labelled as *absent* for the picked 5 classes. Our final balanced dataset consists of 475 images.

In order to carry out our experiments, we implemented two CBIR systems that have been evaluated on the dataset obtained as previously described: a feature-level matching system and a local feature-based bag-of-words pipeline. Both systems rely on SIFT [14] for local feature extraction, which has been extensively used in literature for retrieving images of the same objects from different viewpoints [15], [16], [17]. For the sake of compactness, from now on we will refer to the first method as SIFT and to second method as BOW.

Following [14], our SIFT system matches features according to minimum Euclidean distance. Moreover, a query feature is matched to a dataset feature only if their distance - multiplied by a threshold - is less than the distance between the query and all the other database features. The obtained matching are then refined by checking for spatial consistency using RANSAC [18]. The implementation of our BOW system follows the works of [15] and [16]. We tried

[1] http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/.

Fig. 2. Examples from the 5 query classes we chose to build our dataset

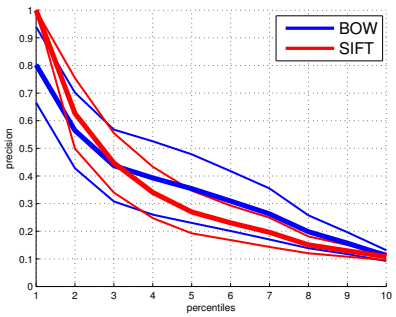
different vocabulary sizes and we found that 50 was the best performer, thus it has been used for the presented experiments. Moreover, 3% most and least frequent visual words are clipped from the vocabulary and not used for image representation and we applied the commonly used *tf-idf* weighting [15].

For each method, PR confidence bands were computed using all query classes and $N_P = 10$ quantiles, according to (2). ANOVA was computed for the APs obtained for each query class and method, resulting in $5 \times 2 = 10$ ANOVA groups, with 5 samples each. All statistics were computed at a significance $\alpha = 0.05$.

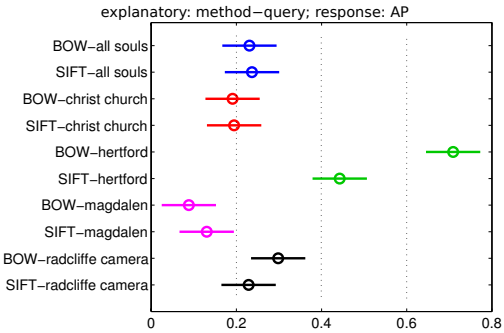
4 Results and Discussion

Computing the traditional AP scores for the two methods under test, we obtain a value of 0.25 for SIFT and 0.30 for BOW. This would suggest that BOW globally outperforms SIFT on this test set. However, the confidence bands obtained for the PR curves of the two methods (Fig. 3(a)) show that, in both cases, the performances are notably variable and the bands consistently overlap. Therefore, we cannot find statistical evidence of the difference between the performances, and even though the AP score is favourable to BOW, it does not necessarily imply that this method is to be preferred for every query class.

Further evidences are brought by the ANOVA multiple comparison experiment, whose outcome is illustrated in Figure 3(b). The figure represents the confidence intervals for the different method-query class factors. As a general comment, SIFT seems more stable showing a slightly smaller variance across the query set. Considering differences across queries, the test does not find a significant difference between the methods for 4 out of 5 classes. The intervals for the classes *all souls* and *christ church* are completely overlapped so it is not possible to make considerations in favour of either one or the other method. Concerning *magdalen* class, we cannot observe a significantly best performance, from a statistical point of view, even if SIFT seem to be slightly preferable. Visually, this class does not particularly differ from *all souls* and *christ church*, sharing with them many local recurring patterns. We suspect that spatial consistency played an important role in discriminating this class from the others, and it determined the success of SIFT method. On the other hand, BOW is significantly better in dealing with *hertford* class, which is the best case for both methods. This class collects images of a building whose structure is sensibly different from the



(a) PR curves and corresponding confidence bands. The curves are computed for the whole query set, i. e. they show the variability across all the 5 classes. For SIFT and BOW the AP scores obtained are respectively 0.25 and 0.30.



(b) ANOVA multiple comparison outcome. Explanatory variables are method-class pairs. The response varible is the AP score. If two intervals are disjoint, two estimates are significantly different. They are not significantly different if the intervals overlap.

Fig. 3. Results from the performed experiments

buildings of other classes. So, we might argue that the presence of very distinguishable features made the task easier for the algorithms, especially favouring the generalization properties of the BOW approach. This consideration can be extended to the *radcliffe camera* class. Even though the test outcome has no statistical significance we can practically observe an important difference between the estimated mean values.

5 Conclusion

In this paper we present a study of a new evaluation framework for a better understanding of the performance scores in image retrieval. This is particularly useful when different query classes can be found in the dataset, such as in the case of the Oxford 5k dataset, or in Greek pottery datasets. We proposed the usage of statistical tools in order to estimate the performance variability, both overall and with respect to the different query classes. This variability, usually neglected by the traditional performance metrics (e. g. mAP score), can reflect the method robustness and allows for a more informed comparison between methods, especially when the query set is particularly heterogeneous.

A main concern for the proposed approach is the number of samples (individuals) for each ANOVA factor, which, being as low as in the current case, it drops ANOVA discriminative power. This implies that less difference might be detected, even though it was possible to observe important differences between the performances for some query classes. This validates our variability study and encourages searching for alternative statistical tools. In particular we plan

to apply mixed model with random effects [19] to increase the discriminative power. Such models are more flexible than ANOVA and allow to to identify explanatory variables for complex designs.

Acknowledgments. Work supported by Spanish projects TIN2012-33116 and TIN2012-37475-C02-02.

References

1. Everingham, M., Ali Eslami, S.M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: Assessing the significance of performance differences on the pascal voc challenges via bootstrapping. Tech. rep. (2013)
2. Bertail, P., Clemencon, S., Vayatis, N.: On bootstrapping the roc curve. In: NIPS, pp. 137–144. Curran Associates Inc. (2008)
3. Clémençon, S., Vayatis, N.: Nonparametric estimation of the precision-recall curve. In: ICML, pp. 185–192 (2009)
4. Macskassy, S.A., Provost, F.J.: Confidence bands for roc curves: Methods and an empirical study. In: ROCAI, pp. 61–70 (2004)
5. Brughi, F., Gil, D., Ramos Terrades, O.: Artistic heritage motive retrieval: an explorative study. Tech. rep. (2013)
6. Crowley, E.J., Zisserman, A.: Of gods and goats: Weakly supervised learning of figurative art. In: BMVC (2013)
7. Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psy. **12**, 387–415 (1975)
8. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics **44**, 837–845 (1988)
9. Wieand, S., Gail, M.H., James, B.R., James, K.L.: A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika **76**(3), 585–592 (1989)
10. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
11. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: ICML, pp. 233–240 (2006)
12. Casella, G., Berger, R.: Statistical inference. Duxbury Press (1990)
13. Hochberg, Y., Tamhane, A.C.: Multiple Comparison Procedures. John Wiley & Sons Inc. (1987)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
15. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV, pp. 1470–1477 (2003)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR, pp. 1–8 (2008)
18. Lebeda, K., Matas, J., Chum, O.: Fixing the locally optimized ransac. In: BMVC, pp. 1–11 (2012)
19. Badiella, L., Puig, P., Leton, E.: Evaluacion diagnostica mediante curvas roc. Tech. rep. (2010)