The University of Surrey Visual Concept Detection System at ImageCLEF@ICPR: Working Notes

M. A. Tahir, F. Yan, M. Barnard, M. Awais, K. Mikolajczyk, and J. Kittler
Centre for Vision, Speech and Signal Processing,University of Surrey, UK
{*m.tahir,f.yan,mark.barnard,m.rana,k.mikolajczyk,j.kittler*}*@surrey.ac.uk*

Abstract

Visual concept detection is one of the most important tasks in image and video indexing. This paper describes our system in the ImageCLEF@ICPR Visual Concept Detection Task which ranked first for large-scale visual concept detection tasks in terms of Equal Error Rate (EER) and Area under Curve (AUC) and ranked third in terms of hierarchical measure. The presented approach involves state-of-the-art local descriptor computation, vector quantisation via clustering, structured scene or object representation via localised histograms of vector codes, similarity measure for kernel construction and classifier learning. The main novelty is the classifier-level and kernel-level fusion using Kernel Discriminant Analysis with RBF/Power Chi-Squared kernels obtained from various image descriptors. For 32 out of 53 individual concepts, we obtain the best performance of all 12 submissions to this task.

1. Introduction

In the Digital Economy of the future it is expected that large repositories of digital information of various type will be compiled and stored, including documents, images, video, music and voice recordings. Digital images and videos especially will require advanced storage and search technology, commonly referred to as content-based multimedia information retrieval (CBMIR) technology. Visual concept detection (VCD) is one of the most important tasks in CBMIR. It aims at annotating images using a vocabulary defined by a set of concepts of interest including scenes types (mountains, snow etc), objects (plants, car etc), and certain named entities (person, place etc). A standard approach to VCD has been established in the community. This approach involves local descriptor computation, vector quantisation via clustering, structured scene or object representation via localised histograms of vector codes, similarity measure for kernel construction and classifier learning. A significant effort has been invested in searching for better solutions in each of these topics.

Although many promising methods have been proposed in each topic, these methods are often difficult to integrate to work as a robust system because different components are optimised individually and evaluated on different datasets. ImageCLEF@ICPR PhotoAnnotation [7] is an evaluation initiative that aims at comparing image-based approaches in the consumer photo domain. It consists of two main tasks: the visual concept detection and annotation tasks. The aim of this paper is to present our system in the Large-Scale Visual Concept Detection Task which ranked *first* in terms of EER and AUC and ranked *third* in terms of hierarchical measure. For the concepts, an average AUC of 86% could be achieved, including concepts with an AUC as high as 96%. For 32 out of 53 individual concepts, we obtained the best performance of all 12 submissions addressing this task.

The rest of paper is organised as follows. Section 2 describes the system followed by a description of the methods submitted in Section 3. Experiments and the results are discussed in Section 4. Section 5 concludes the paper.

2. Visual Concept Detection System

The visual concept detection problem can be formulated as a two class pattern recognition problem. The original data set is divided into N data sets where $Y = \{1, 2, ..., N\}$ is the finite set of concepts. The task is to learn one binary classifier $h_a : X \rightarrow \{\neg a, a\}$ for each concept $a \in Y$. We may choose various visual feature extraction methods to obtain X. Figure 1 shows the visual concept detection system adopted in this paper. It follows the standard bag-of-words model [8] that has become the method of choice for visual categorisation [11, 9, 12]. The system consists of six main components. Each component is implemented via state-of-the-art techniques. These components are described below.

Sampling Strategy: The model first extracts specific points in an image using a point sampling strategy.

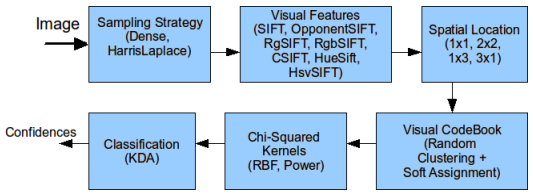

Figure 1. Visual Concept Detection System.

Two methods have been chosen: Dense sampling, and Harris-Laplace. Dense sampling selects points regularly over the image at fixed pixel intervals. Typically, around 10,000 points are sampled per image at an interval of 6 pixels. The Harris-Laplace salient point detector [6] uses the Harris corner detector to find potential feature locations and then selects a subset of these points for which the Laplacian-of-Gaussians reaches a maximum over scale.

Visual Feature Extraction: To describe the area around the sampled points, we use the SIFT descriptor [5], HSV Sift, HUE Sift, two extensions of SIFT [6] and four extensions of SIFT to colour [11]: OpponentSIFT, RGSIFT, C-SIFT, RGB-SIFT. These descriptors have specific invariance properties with respect to common changes in illumination conditions and have been shown to improve visual categorisation accuracy [11].

Spatial Location and Visual Codebook: In order to create a representation for each image we employ the commonly used bag of visual words technique. All the descriptors in the training set are clustered using the kmeans algorithm into 4000 clusters. This is a hierarchical process, first the data is clustered into 10 high level clusters and then 400 lower level clusters. A histogram is then produced for each image in the training set. This 4000 bin histogram is populated using the *Codeword Uncertainty* method presented by Van Gemert et al [3] where the histogram entry of each visual codeword w is given by

$$UNC(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{K_\sigma(D(w,r_i))}{\sum_{j=1}^{|V|}K_\sigma(D(w_j,r_i))}, \quad (1)$$

where n is the number of descriptors in the image, $D(w,r_i)$ is the Euclidean distance between the descriptor r_i and its cluster centre on codeword w, K is a Gaussian kernel with smoothing factor σ and V is the visual vocabulary containing the codeword W. This method of histogram generation has been shown to perform well in the visual concept detection [11].

Classification using Kernel Discriminant Analysis and Spectral Regression: Kernel based learning methods are commonly regarded as a solid choice in order to learn robust concept detectors from large-scale visual codebooks. In recent work [9], we have successfully used kernel discriminant analysis using spectral regression (SRKDA), initially introduced by Cai et al [2], for large-scale image and video classification problems. This method combines the spectral graph analysis and regression for an efficient large matrix decomposition in KDA. It has been demonstrated in [2] that it can achieve an order of magnitude speedup over the eigen-decomposition while producing smaller error rate compared to state-of-the-art classifiers. Later in [9], we have shown the effectiveness of SRKDA for large scale concept detection problems. In addition to superior classification results when compared to existing approaches, it can provide an order of magnitude speed-up over support vector machine. The main computationally intensive operation is Cholesky decomposition, which is actually independent of the number of labels. For more details please refer to [9].

The total computational cost of SRKDA for all concepts in visual concept detection is $\frac{1}{6}m^3 + m^2Nc$ flams where flam is a compound operation consisting of one addition and one multiplication and m is the number of samples. Compared to the cost of ordinary KDA for VCD, $(N \times (\frac{9}{2}m^3 + m^2c))$ flams, SRKDA achieves an order of magnitude ($27N$ times) speed-up over KDA which is massive for large scale image/video datasets.

3. Submitted Runs

We have submitted five different runs described below. All runs use 72 kernels generated from different visual feature representations (2 sampling strategies, 9 different descriptor types and 4 spatial location grids). In this paper, we use only visual information. Future research includes usage of EXIF metadata provided for the photos. The main novelty is the classifier-level and kernel-level fusion using SRKDA with RBF/Power Chi-Squared kernels obtained from various image descriptors. It is worth mentioning that we have also evaluated the performance using SVM with the same kernels and based on the results from validation set, KDA is superior to SVM. These runs are described below:

RUN1: Classifier-level Fusion using RBF Kernels (CLF-KDA) In general, the discriminatory power of kernel classifiers comes directly from the complexity of the underlying kernels. In this run, we have used standard RBF kernel with Chi-squared distance metric: $k(\vec{F},\vec{F'}) = e^{-\frac{1}{A}dist_{\chi^2}(\vec{F},\vec{F'})}$ where A is a scalar which normalises the distances. Following [12], A is set to the average χ^2 distance between all elements of the kernel matrix. Each kernel is then trained using SR-KDA with the regularization parameter, δ, tuned using the validation set. The output from each classifier is then com-

bined using the AVG rule [4]. It is worth noting that for this run we have tried various combination rules such as MAX, MIN, MEDIAN. The best result on the validation set is obtained by the AVG rule and is reported here.

RUN2: Kernel-level Fusion using RBF Kernels (KLF-KDA) In this run, the same RBF kernels with χ^2 distance as in RUN1 are used. However, instead of classifier level fusion, this run uses kernel level fusion with uniform weighting. This corresponds to taking the Cartesian product of the features spaces of the base kernels. Once the kernels are combined, kernel Fisher discriminant analysis is applied as the classifier.

RUN3: Stacked KDA This run uses the classifier in RUN2 as a base classifier for each of the 53 concepts to produce 53 scores. These scores are used as feature vectors and another RBF kernel is built with these features. Note however, for some concepts, not all 53 scores are used for building this kernel. In cases where we have information about the correlation of the concepts, for example, for the disjoint concepts "single person", "small group", "big group", and "no persons", only the scores of the base classifiers for these 4 concepts are used. The new kernel is then added to the set of kernels and kernel FDA classifiers are trained in a second round.

RUN4: Classifier-level fusion using Power Kernels (CLF-KDA-Power) Conditional positive definite kernels have also drawn attention during the last decade and proved successful in image recognition using SVM [1]. In recent work [10], we have modified SRKDA to support conditional positive definite kernels such as power kernels. The main idea is to use LDL^T decomposition instead of Cholesky decomposition. For more details, please refer to [10]. In this run, we have used Power kernel with Chi-squared distance metric: $k(\vec{F}, \vec{F'}) = -dist_{\chi^2}(\vec{F}, \vec{F'})^\beta$ (Conditional Positive Definite if $0 < \beta < 2$). Each power kernel is then trained using modified SRKDA with the regularization parameters δ and β tuned using the validation set. The output from each classifier is then combined using the AVG rule.

RUN5: Based on the performance on the validation set, this run selects the best of RUN2 and RUN3 for each concept.

4. Experimental Results

The ImageCLEF@ICPR dataset consists of 18000 images of 53 different object classes such as animals, vehicles, etc. The dataset is divided into a predefined "trainval" set (8000 images) and "test" set (10000 images). The "trainval" dataset is further divided for validation purpose into a training set containing 5000 images and a validation set containing 3000 images. The ground truth for the test sets is not released to avoid over-fitting of classifiers.

The Equal Error Rate (EER) and the Area under Curve (AUC) are used as measures for large-scale visual concept detection while an hierarchical measure is used to provide a score for the annotation performance for each image.

Results on Validation Set: We first evaluate the classifiers performance on the validation set using different techniques and then compare it to the state-of-the art systems that produced the top results in ImageCLEF@ICPR Challenge. Table 1 shows the performance of our runs including the best and worst descriptors. It is clear from the table that fusion of information either at classifier-level or kernel-level has significantly improved the performance. It is interesting to observe that while RBF-CLF has the best performance both in terms of mean AUC and EER, this run ranked top in only few concepts when compared to other submitted runs. Further, it should be noted that we have also tried to select the best combination of descriptors using search techniques such as Sequential Forward Search but were unable to get any improvement at all on the validation set. Since all of the classifiers contain complementary information, we have used all 9 descriptors with four spatial locations and 2 sampling strategies in our experiments.

Table 1. Comparison of different runs on ImageCLEF@ICPR Validation Set. Ind. Best Descriptor = DS-SIFT-1x1 for AUC, HS-SIFT-2x2 for EER. Ind. Worst Descriptor = DS-HSVSIFT-1x1 for AUC, DS-HSVSIFT-3x1 for EER.

Method	AUC	#WINs	EER	#WINs
Ind. Best	0.7843	-	0.2811	-
Ind. Worst	0.7347	-	0.3236	-
CLF-KDA	0.8424	5	0.2319	10
CLF-KDA-Power	0.8379	12	0.2348	15
KLF-KDA	0.8423	23	0.2319	13
Stacked KDA	0.8400	13	0.2324	15

Results on Test Set: Table 2 shows the performance of best run of each team evaluated independently by the organizers. The best performance using EER and AUC is achieved by our method based on classifier-level fusion using RBF Kernels. In fact the top 2 methods are clearly significantly better than all the other methods. Table 2 also shows the performance using the hierarchical measure in which our method (RUN5) ranked *third*. Technical details of the approaches by other groups have not been published but from the previous workshop on ImageCLEF 2009 [7], ISIS approach is an extension of the system proposed in [11] where SIFT features are extracted in different colour spaces. The learning step is based on SVM with χ^2 kernel which differs from our system mainly where RBF/Power kernels with KDA is

used in the classification stage. For 32 out of 53 individual concepts, we obtain the best performance of all submissions to this task when AUC is used as the evaluation criterion; more than twice when compared with second best method. For EER, the best performance is obtained in 29 out of the 53 individual concepts. These results clearly show the effectiveness of our system for large-scale visual concept detection.

Table 2. The team runs of ImageCLEF@ICPR Photo Annotation Task (from the official evaluations). HM = Hierarchical measure.

Group	EER	#WINs	AUC	#WINs	HM
CVSSP	**0.2136**	29	**0.8600**	32	0.6900
ISIS	0.2182	17	0.8568	15	**0.7836**
IJS	0.2425	5	0.8321	3	0.7065
CNRS	0.2748	1	0.7927	2	0.4204
AVEIR	0.2848	0	0.7848	1	0.5602
MMIS	0.3049	0	0.7566	0	0.5027
LSIS	0.3106	0	0.7490	0	0.5067
UPMC/LIP6	0.3377	0	0.7159	0	0.4034
ITI	0.3656	1	0.5917	0	0.4023
MRIM	0.3831	0	0.6393	0	0.5801
TRS2008	0.4152	0	0.6200	0	0.3270
UAIC	0.4762	0	0.1408	0	0.6781

Table 3 shows the performance of our runs in terms of AUC on a few individual concepts. It is observed that the performance may vary in different concepts. The results indicate that RBF kernels perform quite well when class imbalance is not severe (for example in Day, No-Blur etc). On the other hand, in many highly unbalanced categories like Desert, Lake etc., Power Kernel performs quite well. In some concepts, stacking also has significant effect on the performance e.g. Fancy approx. a 4% improvement over the best run. It is observed that fusion at decision-level or feature-level yields very similar performance on this dataset with the results showing slightly in favour of the classifier-level fusion both in terms of EER and AUC. But the kernel-level fusion has speed advantage over the classifier-level fusion as only one classifier is required to train while the classifier-level fusion requires separate classifiers for the individual descriptors. The results also indicate that RBF-CLF (RUN1) ranked top in the majority of the concepts over other runs indicating that other runs may have overfitted during parameter optimization on the validation set. For RBF-CLF, the same regularisation parameter, $\delta = 0.1$, is used for all concepts while for RBF-KLF/Stacking, δ is tuned for every concept. Similarly, for power kernel, β is also tuned along with δ on the validation set.

5. Conclusions

Our focus on machine learning methods for concept detection in ImageCLEF@ICPR has proved successful. Our method ranked top for the large-scale visual concept detection task in terms of both EER and AUC. For 32 out of 53 individual concepts, we obtained the best performance of all submissions addressing this task. The main novelty is the use of classifier-level and kernel-level fusion with Kernel Discriminant Analysis employing RBF/Power Chi-Squared kernels obtained from various image descriptors. Future work aims to combine ontology (hierarchy and relations) with visual information to improve the performance.

Acknowledgements: This work was supported by EU Vidi-Video project.

Table 3. Comparison of AUC for some individual concepts in ImageCLEF@ICPR Test Set. GT = Ground Truth.

Concept	GT	RUN1	RUN2	RUN3	RUN4
Desert	31	0.8752	0.8762	0.8689	**0.8977**
Lake	90	0.8991	0.8959	0.9015	**0.9122**
Snow	128	**0.8925**	0.8846	0.8773	0.8819
Fancy	1174	0.5881	0.5839	**0.6100**	0.6051
Single-Person	1701	0.8184	0.8192	**0.8342**	0.8019
Sky	1977	0.9582	0.9582	**0.9587**	0.9475
Day	4313	**0.8660**	0.8656	0.8600	0.8509
No-Blur	5274	**0.8578**	0.8573	0.8562	0.8432
Mean		**0.8600**	0.8588	0.8534	0.8547
#WINs		21	12	6	14

References

[1] S. Boughorbel, J. P. Tarel, and N.Boujemaa. Conditionally positive definite kernels for SVM based image recognition. In *Proc. of ICME*, Amsterdam, The Netherlands, 2005.

[2] D. Cai, X. He, and J. Han. Efficient kernel discriminat analysis via spectral regression. In *In Proc. of the ICDM*, 2007.

[3] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *PAMI*, in press, 2009.

[4] J. Kittler, M. Hatef, Robert P. W. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.

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

[6] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[7] S. Nowak and P. Dunker. Overview of the clef 2009 large scale visual concept detection and annotation task. In *CLEF working notes*, Corfu, Greece, 2009.

[8] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proc. of the ICCV*, 2003.

[9] M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K.E.A. van de Sande, and T. Gevers. Visual category recognition using spectral regression and kernel discriminant analysis. In *Proc. of the 2nd International Workshop on Subspace 2009, In Conjunction with ICCV 2009*, Kyota, Japan, 2009.

[10] M. A. Tahir, J. Kittler, F. Yan, and K. Mikolajczyk. Kernel discriminant analysis using triangular kernel for semantic scene classification. In *Proc. of the 7th International Workshop on CBMI*, Crete, Greece, 2009.

[11] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, in press, 2010.

[12] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007.