

Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)

Held in conjunction with [PASCAL Visual Object Classes Challenge 2012 \(VOC2012\)](#).
[Back to Main page](#)

All results

- [Task 1 \(classification\)](#)
- [Task 2 \(localization\)](#)
- [Task 3 \(fine-grained classification\)](#)
- [Team information and abstracts](#)

Task 1

Team name	Filename	Error (5 guesses)	Description
SuperVision	test-preds-141-146.2009-131-137-145-146.2011-145f.	0.15315	Using extra training data from ImageNet Fall 2011 release
SuperVision	test-preds-131-137-145-135-145f.txt	0.16422	Using only supplied training data
ISI	pred_FVs_wLACs_weighted.txt	0.26172	Weighted sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.
ISI	pred_FVs_weighted.txt	0.26602	Weighted sum of scores from classifiers using each FV.
ISI	pred_FVs_summed.txt	0.26646	Naive sum of scores from classifiers using each FV.
ISI	pred_FVs_wLACs_summed.txt	0.26952	Naive sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.

OXFORD_VGG	test_adhocmix_classification.txt	0.26979	Mixed selection from High-Level SVM scores and Baseline Scores, decision is performed by looking at the validation performance
XRCE/INRIA	res_1M_svm.txt	0.27058	
OXFORD_VGG	test_finecls_classification.txt	0.27079	High-Level SVM over Fine Level Classification score, DPM score and Baseline Classification scores (Fisher Vectors over Dense SIFT and Color Statistics)
OXFORD_VGG	test_baseline_classification.txt	0.27302	Baseline: SVM trained on Fisher Vectors over Dense SIFT and Color Statistics
University of Amsterdam	final-UvA-Isvoc2012test.results.val	0.29576	See text above
XRCE/INRIA	res_64k_svm.txt	0.33419	
LEAR-XRCE	submit_i12_d0512_mix.txt	0.34464	Trained on ILSVRC'12 - using a mixture of NCM classifiers
LEAR-XRCE	submit_i12_d0512_k1.txt	0.36184	Trained on ILSVRC'12 - using NCM
LEAR-XRCE	submit_i10_d0512_mix.txt	0.38006	Trained on ILSVRC'10 - using a mixture of NCM classifiers
LEAR-XRCE	submit_i10_d0512_k1.txt	0.41048	Trained on ILSVRC'10 - using NCM

Task 2

Team name	Filename	Error (5 guesses)	Description
-----------	----------	-------------------	-------------

SuperVision	test-rect-preds-144-cloc-141-146.2009-131-137-145-	0.335463	Using extra training data for classification from ImageNet Fall 2011 release
SuperVision	test-rect-preds-144-cloc-131-137-145-135-145f.txt	0.341905	Using only supplied training data
OXFORD_VGG	test_adhocmix_detection.txt	0.500342	Re-ranked DPM detection over Mixed selection from High-Level SVM scores and Baseline Scores, decision is performed by looking at the validation performance
OXFORD_VGG	test_finecls_detection_best_bbox.txt	0.50139	Re-ranked DPM detection over High-Level SVM Scores
OXFORD_VGG	test_finecls_detection_first_bbox.txt	0.522189	Re-ranked DPM detection over High-Level SVM Scores - First bbox selection heuristic
OXFORD_VGG	test_baseline_detection.txt	0.529482	DPM detection over baseline classification scores
ISI	result2.txt	0.536474	We use the cascade object detection with deformable part models, restricting the sizes of bounding boxes.
ISI	result.txt	0.536546	We use the cascade object detection with deformable part models, restricting the sizes of bounding boxes.

Task 3

Team name	Filename	mAP	Description

ISI	CSIFT_GIST_RGBSIFT.dat	0.322524	We represent images by Fisher Vectors computed respectively from CSIFT, GIST, RGBSIFT. We train linear classifiers on each FV by Passive-Aggressive algorithm. Scores are summed up for prediction.
ISI	CSIFT_GIST_LBP_RGBSIFT.dat	0.319673	We represent images by Fisher Vectors computed respectively from CSIFT, GIST, LBP, RGBSIFT. We train linear classifiers on each FV by Passive-Aggressive algorithm. Scores are summed up for prediction.
ISI	GIST_LBP_RGBSIFT.dat	0.315712	We represent images by Fisher Vectors computed respectively from GIST, LBP, RGBSIFT. We train linear classifiers on each FV by Passive-Aggressive algorithm. Scores are summed up for prediction.
XRCE/INRIA	res_1M_svm_nocrop.dat	0.309932	
Uni Jena	scores.dat	0.245897	

Team information and abstracts

Team name	Team members	Abstract
ISI	Naoyuki Gunji (the Univ. of Tokyo), Takayuki Higuchi (the Univ. of Tokyo), Koki Yasumoto (the Univ. of Tokyo), Hiroshi Muraoka (the Univ. of Tokyo), Yoshitaka Ushiku (the Univ. of	<p>Task 1: Classification</p> <p>We use multi-class online learning and late fusion techniques with multiple image features.</p> <p>We extract conventional Fisher Vectors (FV) [Sanchez et al., CVPR 2011] and streamlined version of Graphical Gaussian Vectors (GGV) [Harada, NIPS 2012]. For extraction, we use</p>

	<p>Tokyo), Tatsuya Harada (the Univ. of Tokyo & JST PRESTO), Yasuo Kuniyoshi (the Univ. of Tokyo)</p>	<p>not only common SIFT and CSIFT, but also LBP and GIST in a dense-sampling manner.</p> <p>We train linear classifiers using Passive-Aggressive (PA) algorithm [Crammer et al., JMLR 2006].</p> <p>Then we investigate two strategies to combine scores from each feature's classifier. One is to sum all scores simply, and the other is to train another version of PA using the scores. We train the weight for each feature and sum the scores using the weight.</p> <p>Task 2: Classification with localization</p> <p>We extract HOG descriptors from each sliding window.</p> <p>We use the cascade object detection with deformable part models [Felzenszwalb et al., CVPR 2010], restricting the sizes of bounding boxes.</p> <p>We also restrict the candidates of objects for each input image using the predictions of the Task 1.</p> <p>Task 3: Fine-grained classification</p> <p>We represent images using FVs computed from a variety of descriptors. Each descriptors are extracted more densely than those in the Task 1.</p> <p>We train linear classifiers on each FV using PA. Then scores are summed up for prediction.</p>
<p>LEAR-XRCE</p>	<p>Thomas Mensink, LEAR - INRIA Grenoble and TVPA - Xerox Research Centre Europe Jakob Verbeek, LEAR - INRIA Grenoble Florent Perronnin, TVPA - Xerox Research Centre Europe Gabriela Csurka, TVPA - Xerox Research Centre Europe</p>	<p>In our submission we evaluate the performance of the Nearest Mean Classifier (NCM) in the ILSVRC 2012 Challenge. The idea of the NCM classifier is to classify an image to the class with the nearest class-mean. To obtain competitive performance we learn a low rank Mahalanobis distance function, $M = W^T W$, by maximizing the log-likelihood of correct prediction [Mensink et al., ECCV'12].</p> <p>We submit two runs (a and b) where the metric has been learned and parameters has been validated on the ILSVRC'10 training and evaluation set. There is no training on the ILSVRC'12 dataset, except that we had to compute the class means on the ILSVRC'12 training set.</p>

		<p>The other two runs (c and d) use a metric which has been learned on the ILSVRC'12 dataset.</p> <p>Run b differs from run a (and similar d differs from c), in that we use a non-linear extension of the NCM classifier, where each class is represented by k centroids in stead of only a single mean [Mensink et al., TechReport'12]. The metric we use is learned and validated for k=1. For the final classification results we use a mixture of k = [1 5 10 15 20] centroids, where each mixture component has equal weight.</p> <p>Images are represented by Fisher Vectors on SIFT and Local Color Features [Lowe, IJCV'04 and Perronnin et al., ECCV'10], which are early fused into a 64K dimensional feature vector, these vectors are compressed using Product Quantization [Jégou et al., PAMI'11].</p>
<p>OXFORD_VGG</p>	<p>Karen Simonyan, University of Oxford Yusuf Aytar, University of Oxford Andrea Vedaldi, University of Oxford Andrew Zisserman, University of Oxford</p>	<p>In this submission, image classification was performed using a conventional pipeline based on Fisher vector image representation and one-vs-rest linear SVM classifiers. In more detail, two types of local patch features were densely extracted over multiple scales: SIFT and colour statistics. The features were then augmented with patch spatial coordinates and aggregated into two Fisher vectors corresponding to the two feature types. Fisher vectors were computed using GMMs with 1024 Gaussians, resulting in 135K-dimensional representations. To obtain a single feature vector per image, the two Fisher vectors were then stacked. We did not use spatial pyramid representation. To be able to deal with large amounts of training data, product quantisation was employed to compress the image features. Finally, an ensemble of one-vs-rest linear SVMs was trained over stacked features using stochastic sub-gradient method (Pegasos).</p> <p>Localization is performed using DPM (discriminatively trained part based models) detectors (without parts) trained for each class individually. The DPM detectors are boosted via harvesting more bounding boxes from the training set using a semi-supervised approach. Using the validation set the top 5000 images for each class are shortlisted via image classification score and then detection is performed using DPMs. After that, for each class individually another boundingbox-aware classification model is trained from the cropped images using the max-scored bounding box for each image. In this fine-level bounding box classification, we used features similar to those used for image classification (stacked dense SIFT and colour Fisher vectors). Finally, for</p>

		<p>each class, a high-level SVM is trained over the image classification score, DPM max-detection score and the score from fine-level bounding box classification. Using the scores from the high-level SVM, top 5000 shortlist from the test set is re-ranked.</p>
<p>SuperVision</p>	<p>Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton</p> <p>University of Toronto</p>	<p>Our model is a large, deep convolutional neural network trained on raw RGB pixel values. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three globally-connected layers with a final 1000-way softmax. It was trained on two NVIDIA GPUs for about a week. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of convolutional nets. To reduce overfitting in the globally-connected layers we employed hidden-unit "dropout", a recently-developed regularization method that proved to be very effective.</p>
<p>Uni Jena</p>	<p>Christoph Göring, Computer Vision Group, Friedrich Schiller University Jena, Germany</p> <p>Erik Rodner, ICSI Vision Group, UC Berkeley, California</p> <p>Alexander Freytag, Uni Jena, Computer Vision Group, Friedrich Schiller University Jena, Germany</p> <p>Joachim Denzler, Uni Jena, Computer Vision Group, Friedrich Schiller University Jena, Germany</p>	<p>Our team tackled task 3 of ILSVRC 2012 challenge - namely fine-grained object classification. We built a final classification system relying on three key ingredients: (1) the combination of different feature types to capture different aspects of objects, namely shape, color and texture, (2) a simple yet efficient part detector together with background elimination using a segmentation approach without any user interaction, and (3) a linear classifier with an efficient kernel approximation to ensure computation times within a few hours even for this large-scale dataset. Details for every step follow in the subsequent paragraphs.</p> <p>For differentiating between hundreds of dog categories, not few but many details matter. Therefore, we represent images by a combination of different sources of information. Shape of objects is captured using a bag-of-words histogram of opponent sift descriptors which are densely sampled from the image. In addition, we extract color information using colorname histograms. Finally, we compute local binary patterns to capture texture information, which is helpful for differentiating between different fur structures. We add spatial information to every type of feature by extracting not a single feature per image but a pyramid histograms representation.</p> <p>Following state-of-the-art approaches, we additionally extract part based information. Since bodies of dogs are highly deformable, the parts being most reliably detectable</p>

are their heads. Unfortunately, there is no annotation for these parts available in the data and we can not train a standard detector. Therefore, we use a simple head detector by applying a hough circle transform to find eyes and noses and then search for 3 circles that compose a triangle. With this approach we are able to find a large fraction of dog heads in the images. Our detection approach does not work with dark fur, bad illumination conditions, and when the head is not in the picture. Detection results are used to extract an additional sift bag-of-words descriptor from the head region.

Background clutter present in the images might interfere classification. We therefore apply grabcut to all images to consider relevant foreground regions only. For grabcut, a background color model was trained on the pixels outside of the provided bounding box, whereas a foreground color model was trained on pixels inside the bounding box. This initial bounding box segmentation is then refined using iterated graph cuts.

Images are finally represented by a combination of all previously described features.

For classification, we use the liblinear svm with a one-vs-all multiclass approach. Due to the linearity of the classifier, computing classification scores is extremely fast which makes it feasible for this large-scale dataset. However, the gain of speed has the drawback of a diminished discriminative power. We overcome this drawback by utilizing homogenous kernel maps to approximate a chi2-kernel. With this combination we are able to combine the speed of a linear svm with the discriminative power of kernel-based methods. With all details mentioned, liblinear is able to train a model using the 20,500 training examples in less than 4 hours using 70Gb RAM.

University of Amsterdam

Koen E. A. van de Sande
Amir Habibian
Cees G. M. Snoek

We extend the fine grained coding approach of last years LSVRC classification winners [1] by fine grained color descriptors and a calibrated SVM with a cutting plane solver. We provide the cutting plane solver 10% of the negative examples per class to train an exact model (not stochastic). For each fine grained color descriptors we train a separate SVM which is fused at the classifier level. This is more precise and more efficient than training on a concatenated version of the features. Last years UvA system used Platt's sigmoid to calibrate classifier scores, which involves expensive 5-fold cross-validation. Therefore, this year we used an unsupervised calibrator based on extreme value

		<p>theory [2]. It fits a Weibull to the classifier scores which is subsequently used for score normalization. Overall, the error rate is approximately 8% lower than last years UvA system on the validation set.</p> <p>[1] F. Perronnin and J. Sanchez, Compressed Fisher vectors for Large Scale Visual Recognition, Large Scale Visual Recognition workshop, ICCV 2011.</p> <p>[2] W. Scheirer, N. Kumar, P. Belhumeur and T. Boult, Multi-Attribute Spaces: Calibration for Attribute Fusion and Similarity Search, CVPR 2012</p>
XRCE/INRIA	<p>Florent Perronnin, XRCE Zeynep Akata, XRCE/INRIA Zaid Harchaoui, INRIA Cordelia Schmid, INRIA</p>	<p>Our low-level descriptors are the SIFT of [Lowe, IJCV 2004] and the color features of [Perronnin et al., ECCV 2010]. They are aggregated into image-level features using the Fisher Vector (FV) of [Perronnin et al, ECCV 2010]. Because of the high-dim of the FVs, they are compressed with Product Quantization (PQ) as proposed in [Sanchez and Perronnin, CVPR 2011].</p> <p>We train linear SVMs in a one-vs-rest manner using the good practices of [Perronnin et al., CVPR 2012]. This leads on the validation set to approx. 2% improvement with respect to last year's training strategy.</p> <p>For task 1 we submitted two systems:</p> <ul style="list-style-type: none"> - res_64k_svm.txt: we use 64K FVs obtained by concatenating 32K-dim SIFT and color FVs. - res_1M_svm.txt: 0.5M-dim SIFT and color FVs are computed and classified separately. The SIFT and color results are merged with late fusion (weighted averaging). <p>For task 3 we report results only with the 0.5M-dim SIFT and color FVs and late fusion. We submitted two systems:</p> <ul style="list-style-type: none"> - res_1M_svm_nocrop.dat: without using bounding boxes - res_1M_svm_crop.dat: with bounding boxes