# Summary of References Related to Convolutional Neural Networks

Miquel Perello, E-mail: miquel.perellonieto@aalto.fi

May 14, 2014

## Contents

# 1 Introduction

# 2 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex [19]

## 2.1 Original Abstract

*None*

# 3 Receptive fields and functional architecture of monkey striate cortex [20]

## 3.1 Original Abstract

*1. The striate cortex was studied in lightly anaesthetized macaque and spider monkeys by recording extracellularly from single units and stimulating the retinas with spots or patterns of light. Most cells can be categorized as simple, complex, or hypercomplex, with response properties very similar to those previously described in the cat. On the average, however, receptive fields are smaller, and there is a greater sensitivity to changes in stimulus orientation. A small proportion of the cells are colour coded.2. Evidence is presented for at least two independent systems of columns extending vertically from surface to white matter. Columns of the first type contain cells with common receptive-field orientations. They are similar to the orientation columns described in the cat, but are probably smaller in cross-sectional area. In the second system cells are aggregated into columns according to eye preference. The ocular dominance columns are larger than the orientation columns, and the two sets of boundaries seem to be independent.3. There is a tendency for cells to be grouped according to symmetry of responses to movement; in some regions the cells respond equally well to the two opposite directions of movement of a line, but other regions contain a mixture of cells favouring one direction and*

*cells favouring the other.4. A horizontal organization corresponding to the cortical layering can also be discerned. The upper layers (II and the upper two-thirds of III) contain complex and hypercomplex cells, but simple cells are virtually absent. The cells are mostly binocularly driven. Simple cells are found deep in layer III, and in IV A and IV B. In layer IV B they form a large proportion of the population, whereas complex cells are rare. In layers IV A and IV B one finds units lacking orientation specificity; it is not clear whether these are cell bodies or axons of geniculate cells. In layer IV most cells are driven by one eye only; this layer consists of a mosaic with cells of some regions responding to one eye only, those of other regions responding to the other eye. Layers V and VI contain mostly complex and hypercomplex cells, binocularly driven.5. The cortex is seen as a system organized vertically and horizontally in entirely different ways. In the vertical system (in which cells lying along a vertical line in the cortex have common features) stimulus dimensions such as retinal position, line orientation, ocular dominance, and perhaps directionality of movement, are mapped in sets of superimposed but independent mosaics. The horizontal system segregates cells in layers by hierarchical orders, the lowest orders (simple cells monocularly driven) located in and near layer IV, the higher orders in the upper and lower layers.*

# 4 Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position [12]

## 4.1 Original Abstract

*A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by "learning without a teacher", and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is given a nickname "neocognitron". After completion of self-organization, the network has a structure similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. The network consists of an input layer (photoreceptor array) followed by a cascade connection of a number of modular structures, each of which is composed of two layers of cells connected in a cascade. The first layer of each module con-*

*sists of "S-cells', which show charac- teristics similar to simple cells or lower order hyper- complex cells, and the second layer consists of "C-cells" similar to complex cells or higher order hypercomplex cells. The afferent synapses to each S-cell have plasticity and are modifiable. The network has an ability of unsupervised learning: We do not need any "teacher" during the process of self- organization, and it is only needed to present a set of stimulus patterns repeatedly to the input layer of the network. The network has been simulated on a digital computer. After repetitive presentation of a set of stimulus patterns, each stimulus pattern has become to elicit an output only from one of the C-cells of the last layer, and conversely, this C-cell has become selectively responsive only to that stimulus pattern. That is, none of the C-cells of the last layer responds to more than one stimulus pattern. The response of the C-cells of the last layer is not affected by the pattern's position at all. Neither is it affected by a small change in shape nor in size of the stimulus pattern. 1.*

## 4.2 Main points

- Reiteration of self-organized by "learning without a teacher"
- Similar structure to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel.
- Network structure:
  - Input layer (photoreceptor array)
  - Cascade of modules each one with :
    - S-cells: in the first layer Simple cells or lower order hypercomplex cells
    - C-cells: in the second layer Complex cells or higher order hypercomplex cells
- Hubel and Wiesel : the neural network in the visual cortex has a hierarchy structure:
  - LGB (Lageral Geniculate Body)
  - Simple cells
  - Complex cells

  - Lower order hypercomplex cells
  - Higher order hypercomplex cells

- a cell in a higher stage generally has tendency to respond selectively to a more complicated feature of the stimulus pattern
- we extend the hierarchy model of Hubel and Wiesel, and **hypothesize** the existance of a similar hierarchy structure even in hte stages higher than hypercomplex cells.
- In the last module, the receptive field of each C-cell becomes so large as to cover the whole area of input layer $U_0$, and each C-plane is so determined as to have only one C-cell
- The output of an S-cell in the $k_l$-th S-plane in the l-th module is described below

# 5 Generalization and network design strategies [30]

## 5.1 Original Abstract

*An interestmg property of connectiomst systems is their ability tolearn from examples. Although most recent work in the field concentrateson reducing learning times, the most important feature of a learning ma-chine is its generalization performance. It is usually accepted that goodgeneralization performance on real-world problems cannot be achievedunless some a pnon knowledge about the task is butlt Into the system.Back-propagation networks provide a way of specifymg such knowledgeby imposing constraints both on the architecture of the network and onits weights. In general, such constramts can be considered as particulartransformations of the parameter spaceBuilding a constramed network for image recogmtton appears to be afeasible task. We descnbe a small handwritten digit recogmtion problemand show that, even though the problem is linearly separable, single layernetworks exhibit poor generalizatton performance. Multtlayer constrainednetworks perform very well on this task when orgamzed in a hierarchicalstructure with shift invariant feature detectors.These results confirm the idea that minimizing the number of freeparameters in the network enhances generalization.*

# 6 Backpropagation applied to handwritten zip code recognition [32]

## 6.1 Original Abstract

*The ability of learning networks to generalize can be greatly enhanced by providing constraints from the task domain. This paper demonstrates how such constraints can be integrated into a backpropagation network through the architecture of the network. This approach has been successfully applied to the recognition of handwritten zip code digits provided by the U.S. Postal Service. A single network learns the entire recognition operation, going from the normalized image of the character to the final classification.*

## 6.2 Main points

# 7 Convolutional networks for images, speech, and time series [31]

## 7.1 Original Abstract

*None*

# 8 Gradient-based learning applied to document recognition [33]

## 8.1 Original Abstract

*Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional neural networks, which are specifically designed to deal*

*with the variability of 2D shapes, are shown to outperform all other techniques. Real-life document recognition systems are composed of multiple modules including field extraction, segmentation recognition, and language modeling. A new learning paradigm, called graph transformer networks (GTN), allows such multimodule systems to be trained globally using gradient-based methods so as to minimize an overall performance measure. Two systems for online handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of graph transformer networks. A graph transformer network for reading a bank cheque is also described. It uses convolutional neural network character recognizers combined with global training techniques to provide record accuracy on business and personal cheques. It is deployed commercially and reads several million cheques per day*

## 8.2 Main points

- LeNet-5
- Clarification: In this paper "stride" is not mentioned, but as Krizhevsky2012 et.al. started using it, new implementations of CNN need to define its value.
- Conv: Convolutional layer
- Subs: Subsampling layer (summed * coefficient + bias)
- Full: Fully connected network
- ERBF: Euclidian Radial Basis Function units
  - input 32x32 pixel image (original images are 28x28)
  - Conv1 :
    - 6@28x28 filter 5x5
    - stride 1
    - Connections $= 5 * 5 * 28 * 28 * 6 + 6 * 28 * 28 = 122,304$
    - Train. param. $= 5 * 5 * 6 + 6 = 156$
  - Subs2 :
    - 6@14x14 range 2x2

    - stride 2
    - Connections $= 6 * 28 * 28 + 6 * 14 * 14 = 5,880$
    - Train. param. = coefficient + bias $= 6 + 6 = 156$
- Conv3 :
    - 16@10x10 filter 5x5
    - stride 1
    - Connections $= 6 * 5 * 5 * 10 * 10 * 10 + 10 * 10 * 16 = 151,600$
    - Train. param. $= 5*5*3*6+5*5*4*9+5*5*6*1+16 = 1,516$
    - Note:
    - This layer is not completly connected, see table 1 for specific connections
    - Expected Connections $= 6 * 5 * 5 * 10 * 10 * 16 + 10 * 10 * 16 = 241,600$
    - Expected train. param $= 5 * 5 * 16 * 6 + 16 = 2416$
- Subs4 :
    - 16@5x5 range 2x2
    - stride 2
    - Connections $= 16 * 10 * 10 + 16 * 5 * 5 = 2,000$
    - Train. param. = coefficient + bias $= 16 + 16 = 32$
- Conv5 :
    - 120@1x1 filter 5x5
    - stride 0
    - Connections and train. param. $= 16*5*5*120+120 = 48,120$
- Full6 : 84 Atanh(Sa)
    - Connections and train. param. $= 120 * 84 + 84 = 10,164$
- ERBF7 : 10
    - Connections and train. param. $= 84 * 10 = 840$

# 9 Independent component analysis applied to feature extraction from colour and stereo images. [18]

## 9.1 Original Abstract

*Previous work has shown that independent component analysis (ICA) applied to feature extraction from natural image data yields features resembling Gabor functions and simple-cell receptive fields. This article considers the effects of including chromatic and stereo information. The inclusion of colour leads to features divided into separate red/green, blue/yellow, and bright/dark channels. Stereo image data, on the other hand, leads to binocular receptive fields which are tuned to various disparities. The similarities between these results and the observed properties of simple cells in the primary visual cortex are further evidence for the hypothesis that visual cortical neurons perform some type of redundancy reduction, which was one of the original motivations for ICA in the first place. In addition, ICA provides a principled method for feature extraction from colour and stereo images; such features could be used in image processing operations such as denoising and compression, as well as in pattern recognition.*

## 9.2 Main points

# 10 Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces [21]

## 10.1 Original Abstract

*Olshausen and Field (1996) applied the principle of independence maximization by sparse coding to extract features from natural images. This leads to the emergence of oriented linear filters that have simultaneous localization in space and in frequency, thus resembling Gabor functions and simple cell receptive fields. In this article, we show that the same principle of independence maximization can explain the emergence of phase- and shift-invariant features, similar to those found in complex cells. This new kind of emergence*

*is obtained by maximizing the independence between norms of projections on linear subspaces (instead of the independence of simple linear filter outputs). The norms of the projections on such "independent feature subspaces" then indicate the values of invariant features.*

### 10.2 Main points

## 11 Why color management? [24]

### 11.1 Original Abstract

*It seems that everywhere you look there is some article or discussion about color management. Why all the fuss? Do I need to management my colors? We have been creating colored artifacts for a very long time and I don't think we have needed color management. So why now? Most of these discussions also refer to the ICC. What is that? These and other questions will be answered in a straightforward manner in plain English. Adobe Systems has pioneered the use of desktop computers for color work, and the author has helped Adobe pick its way down conflicting color paths with confusing road signs over the last 10 years.*

### 11.2 Main points

## 12 Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis [45]

### 12.1 Original Abstract

*Neural networks are a powerful technology forclassification of visual inputs arising from documents.However, there is a confusing plethora of different neuralnetwork methods that are used in the literature and inindustry. This paper describes a set of concrete bestpractices that document analysis researchers can use toget good results with neural networks. The mostimportant practice is getting a training set as large aspossible: we expand the training set by adding a newform of distorted data. The next most important practiceis*

*that convolutional neural networks are better suited forvisual document tasks than fully connected networks. Wepropose that a simple "do-it-yourself" implementation ofconvolution with a flexible architecture is suitable formany visual document problems. This simpleconvolutional neural network does not require complexmethods, such as momentum, weight decay, structure-dependent learning rates, averaging layers, tangent prop,or even finely-tuning the architecture. The end result is avery simple yet general architecture which can yieldstate-of-the-art performance for document analysis. Weillustrate our claims on the MNIST set of English digitimages.*

## 12.2 Main points

- Get a training set as large as possible
- No need of complex methods, such as momentum, weight decay, structure-dependent learning rates, averaging layers, tangent prop, or even finely-tuning the architecture
- Increment dataset by:
  - Affine transformations: translations, scaling, homothety, similarity transformation, reflection, rotation, shear mapping, and compositions.
  - Elastic distortions
- In this paper the authors justify the use of elastic deformations on MNIST data corresponding to uncontrolled oscillations of the hand muscles, dampened by inertia.
- They get the best results on MNIST to date with CNN, affine and elastic transformations of the dataset (0.4% error).

# 13 A convolutional neural network approach for objective video quality assessment [5]

## 13.1 Original Abstract

*This paper describes an application of neural networks in the field of objective measurement method designed to automatically assess the perceived quality of*

*digital videos. This challenging issue aims to emulate human judgment and to replace very complex and time consuming subjective quality assessment. Several metrics have been proposed in literature to tackle this issue. They are based on a general framework that combines different stages, each of them addressing complex problems. The ambition of this paper is not to present a global perfect quality metric but rather to focus on an original way to use neural networks in such a framework in the context of reduced reference (RR) quality metric. Especially, we point out the interest of such a tool for combining features and pooling them in order to compute quality scores. The proposed approach solves some problems inherent to objective metrics that should predict subjective quality score obtained using the single stimulus continuous quality evaluation (SSCQE) method. This latter has been adopted by video quality expert group (VQEG) in its recently finalized reduced referenced and no reference (RRNR-TV) test plan. The originality of such approach compared to previous attempts to use neural networks for quality assessment, relies on the use of a convolutional neural network (CNN) that allows a continuous time scoring of the video. Objective features are extracted on a frame-by-frame basis on both the reference and the distorted sequences; they are derived from a perceptual-based representation and integrated along the temporal axis using a time-delay neural network (TDNN). Experiments conducted on different MPEG-2 videos, with bit rates ranging 2-6 Mb/s, show the effectiveness of the proposed approach to get a plausible model of temporal pooling from the human vision system (HVS) point of view. More specifically, a linear correlation criteria, between objective and subjective scoring, up to 0.92 has been obtained on a- - set of typical TV videos*

# 14 A fast learning algorithm for deep belief nets [14]

## 14.1 Original Abstract

*We show how to use "complementary priors" to eliminate the explaining-away effects thatmake inference difficult in densely connected belief nets that have many hidden layers. Using complementary priors, we derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory. The fast, greedy algorithm is used to initialize a slower learning procedure that*

*fine-tunes the weights using a contrastive version of thewake-sleep algorithm. After fine-tuning, a networkwith three hidden layers forms a very good generative model of the joint distribution of handwritten digit images and their labels. This generative model gives better digit classification than the best discriminative learning algorithms. The low-dimensional manifolds on which the digits lie are modeled by long ravines in the free-energy landscape of the top-level associative memory, and it is easy to explore these ravines by using the directed connections to displaywhat the associativememory has in mind.*

# 15 Reducing the dimensionality of data with neural networks [15]

## 15.1 Original Abstract

*High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.*

# 16 Robust object recognition with cortex-like mechanisms. [44]

## 16.1 Original Abstract

*We introduce a new general framework for the recognition of complex visual scenes, which is motivated by biology: We describe a hierarchical system that closely follows the organization of visual cortex and builds an increasingly complex and invariant feature representation by alternating between a template matching and a maximum pooling operation. We demonstrate the strength of the approach on a range of recognition tasks: From invariant single object recognition in clutter to multiclass categorization problems and*

*complex scene understanding tasks that rely on the recognition of both shape-based as well as texture-based objects. Given the biological constraints that the system had to satisfy, the approach performs surprisingly well: It has the capability of learning from only a few training examples and competes with state-of-the-art systems. We also discuss the existence of a universal, redundant dictionary of features that could handle the recognition of most object categories. In addition to its relevance for computer vision, the success of this approach suggests a plausibility proof for a class of feedforward models of object recognition in cortex.*

# 17 To recognize shapes, first learn to generate images [17]

## 17.1 Original Abstract

*The uniformity of the cortical architecture and the ability of functions to move to different areas of cortex following early damage strongly suggest that there is a single basic learning algorithm for extracting underlying structure from richly structured, high-dimensional sensory data. There have been many attempts to design such an algorithm, but until recently they all suffered from serious computational weaknesses. This chapter describes several of the proposed algorithms and shows how they can be combined to produce hybrid methods that work efficiently in networks with many layers and millions of adaptive connections.*

# 18 Deep learning via semi-supervised embedding [50]

## 18.1 Original Abstract

*We show how nonlinear semi-supervised embedding algorithms popular for use with "shallow" learning techniques such as kernel methods can be easily applied to deep multi-layer architectures, either as a regularizer at the output layer, or on each layer of the architecture. Compared to standard supervised backpropagation this can give significant gains. This trick provides a simple alternative to existing approaches to semi-supervised deep learning whilst*

*yielding competitive error rates compared to those methods, and existing shallow semi-supervised techniques.*

# 19 Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations [34]

## 19.1 Original Abstract

*There has been much interest in unsupervised learning of hierarchical generative models such as deep belief networks. Scaling such models to full-sized, high-dimensional images remains a difficult problem. To address this problem, we present the convolutional deep belief network, a hierarchical generative model which scales to realistic image sizes. This model is translation-invariant and supports efficient bottom-up and top-down probabilistic inference. Key to our approach is probabilistic max-pooling, a novel technique which shrinks the representations of higher layers in a probabilistically sound way. Our experiments show that the algorithm learns useful high-level visual features, such as object parts, from unlabeled images of objects and natural scenes. We demonstrate excellent performance on several visual recognition tasks and show that our model can perform hierarchical (bottom-up and top-down) inference over full-sized images.*

## 19.2 Main points

- Probabilistic max-pooling
- Scale DBN to real-sized images
  - Computationally intractable
  - Need invariance in representation
- RBM
  - Binary valued: Independent Bernoulli random variables
  - Real valued: Gaussian with diagonal covariance
  - Training:

        * Stochastic gradient ascent on log-likelihood of training data
        * Contrastive divergence approximation

- Convolutional RBM
    - detection layers: convolving feature maps
    - pooling layers: shrink the representation
        * Block: CxC from bottom layer
        * Max-pooling : minimizes energy subject to only one unit can be active.
    - Sparsity regularization: hidden units have a mean activation close to a small constant

- Convolutional Deep belief network
    - Stacking CRBM on top of one another
    - Training:
        * Gibbs sampling
        * Mean-field (5 iterations in this paper)

# 20 Learning Deep Architectures for AI [2]

## 20.1 Original Abstract

*Theoretical results suggest that in order to learn the kind of complicated functions that can represent high-level abstractions (e.g., in vision, language, and other AI-level tasks), one may need deep architectures. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state-of-the-art in certain areas. This monograph discusses the motivations and principles regarding learning algorithms for deep architectures, in particular those exploiting as building blocks unsupervised learning of single-layer models such as Restricted Boltzmann Machines, used to construct deeper models such as Deep Belief Networks.*

## 20.2 Main points

# 21 Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning [38]

## 21.1 Original Abstract

*In this paper we present a method for learning class-specific features for recognition. Recently a greedy layer-wise procedure was proposed to initialize weights of deep belief networks, by viewing each layer as a separate restricted Boltzmann machine (RBM). We develop the convolutional RBM (C-RBM), a variant of the RBM model in which weights are shared to respect the spatial structure of images. This framework learns a set of features that can generate the images of a specific object class. Our feature extraction model is a four layer hierarchy of alternating filtering and maximum subsampling. We learn feature parameters of the first and third layers viewing them as separate C-RBMs. The outputs of our feature extraction hierarchy are then fed as input to a discriminative classifier. It is experimentally demonstrated that the extracted features are effective for object detection, using them to obtain performance comparable to the state of the art on handwritten digit recognition and pedestrian detection.*

## 21.2 Main points

- New Convolutional Restricted Boltzmann Machine (C-RBM)
- Comparable state-of-the-art on handwritten digit recognition and pedestrian detection
- RBM
  - Probabilistic model
  - hidden variables independent given observerd data
  - Not capture explicitly spacial structure of images
- C-RBM

  - Include spatial locality and weight sharing
  - Favors filters with high response on training images
  - Unsupervised learning using Contrastive Divergence
  - Layerwise training for stacks of RBMs
  - Convolutional connections are employed in a generative Markov Random Field architecture
  - Hidden units divided into K feature maps
  - Convolution problems
    * Boundary units are withinb a smaller number of subwindows compared to the interior pixels
    * middle pixels may contribute to $K_{xy}$ features
    * Separation of boundary variables ($v^b$) from middle variables ($v^m$)
    * Problems sampling from boundary pixels (not have nough features)
    * Over completeness because of K-features
    * Sampling creates images very similar to the original ones
    * Need of more Gibbs sampling steps
    * Their solution is to fix hidden bias terms $c$ during training

- Multilayer C-RBMs
  - Subsampling takes maximum conditional feature probability over non-overlapping subwindows of feature maps
  - Architecture
    * discriminative layer (SVM)
    * max pooling
    * convolution
    * max pooling
    * convolution
    * input
  - On pedestrians also HOG is used in discriminative layer

- MNIST dataset

  - Discriminative layer with RBF kernel
  - 10 one-vs-rest binary SVMs
  - 1st layer 15 feature maps
  - 2nd layer 2x2 non-overlapping subwindos
  - 3rd layer 15 feature maps
  - 4th layer
- Comparison with Large CNN
  - C-RBM is better when training is small
- Pedestrian dataset
  - 1st layer 7x7 15 feature maps
  - 2nd layer 4x4 subsampling
  - 3rd layer 15x5x5 30 feature maps
  - 4th layer 2x2 subsampling
  - + HOG
  - Discriminative layer with linear kernel

# 22 Evaluation of local spatio-temporal features for action recognition [49]

## 22.1 Original Abstract

*Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature and promising recognition results were demonstrated for a number of action classes. The comparison of existing methods, however, is often limited given the different experimental settings used. The purpose of this paper is to evaluate and compare previously proposed space-time features in a common experimental setup. In particular, we consider four different feature detectors and six local feature descriptors and use a standard bag-of-features SVM approach for action recognition. We investigate the performance of these methods on a total of 25 action classes*

*distributed over three datasets with varying difficulty. Among interesting conclusions, we demonstrate that regular sampling of space-time features consistently outperforms all tested space-time interest point detectors for human actions in realistic settings. We also demonstrate a consistent ranking for the majority of methods over different datasets and discuss their advantages and limitations.*

## 22.2 Main points

- Detectors
  - Harris3D
  - Cuboid
  - Hessian
  - Dense sampling
- Descriptors
  - HOG/HOF
  - HOG3D
  - ESURF (extended SURF)
- Datasets
  - KTH actions
    - 6 human action classes
    - walking, jogging, running, boxing, waving and clapping
    - 25 subjects
    - 4 scenarios
    - 2391 video samples
    - `http://www.nada.kth.se/cvap/actions/`
  - UCF sport actions
    - 10 human action classes
    - winging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging, golf swinging and walking
    - 150 video samples

    - http://crcv.ucf.edu/data/UCF_Sports_Action.php
  - Hollywood2 actions
    - 12 action classes
    - answering the hone, driving car, eating, fighting, geting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up.
    - 69 Hollywood movies
    - 1707 video samples
    - http://www.di.ens.fr/~laptev/actions/hollywood2/

# 23 Actions in context [35]

## 23.1 Original Abstract

*This paper exploits the context of natural dynamic scenes for human action recognition in video. Human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate high correlation with particular scene classes. For example, eating often happens in a kitchen while running is more common outdoors. The contribution of this paper is three-fold: (a) we automatically discover relevant scene classes and their correlation with human actions, (b) we show how to learn selected scene classes from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved recognition of both in natural video. We use movie scripts as a means of automatic supervision for training. For selected action classes we identify correlated scene classes in text and then retrieve video samples of actions and scenes for training using script-to-video alignment. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined in a joint scene-action SVM-based classifier. We report experimental results and validate the method on a new large dataset with twelve action classes and ten scene classes acquired from 69 movies.*

# 24 Learning Convolutional Feature Hierarchies for Visual Recognition [23]

## 24.1 Original Abstract

*We propose an unsupervised method for learning multi-stage hierarchies of sparseconvolutional features. While sparse coding has become an increasingly popularmethod for learning visual features, it is most often trained at the patch level.Applying the resulting filters convolutionally results in highly redundant codesbecause overlapping patches are encoded in isolation. By training convolutionallyover large image windows, our method reduces the redudancy between featurevectors at neighboring locations and improves the efficiency of the overall repre-sentation. In addition to a linear decoder that reconstructs the image from sparsefeatures, our method trains an efficient feed-forward encoder that predicts quasi-sparse features from the input. While patch-based training rarely produces any-thing but oriented edge detectors, we show that convolutional training produceshighly diverse filters, including center-surround filters, corner detectors, cross de-tectors, and oriented grating detectors. We show that using these filters in multi-stage convolutional network architecture improves performance on a number ofvisual recognition and detection tasks*

# 25 Tiled convolutional neural networks [37]

## 25.1 Original Abstract

*Convolutional neural networks (CNNs) have been successfully applied to manytasks such as digit and object recognition. Using convolutional (tied) weightssignificantly reduces the number of parameters that have to be learned, and alsoallows translational invariance to be hard-coded into the architecture. In this pa-per, we consider the problem of learning invariances, rather than relying on hard-coding. We propose tiled convolution neural networks (Tiled CNNs), which usea regular "tiled" pattern of tied weights that does not require that adjacent hiddenunits share identical weights, but instead requires only that hidden units k stepsaway from each other to have tied weights. By pooling over neighboring units,this architecture is able to learn complex invariances (such as scale and rotationalinvariance) beyond translational invari-*

*ance. Further, it also enjoys much of CNNs'advantage of having a relatively small number of learned parameters (such as easeof learning and greater scalability). We provide an efficient learning algorithm forTiled CNNs based on Topographic ICA, and show that learning complex invariantfeatures allows us to achieve highly competitive results for both the NORB andCIFAR-10 datasets.*

# 26 Convolutional learning of spatio-temporal features [48]

## 26.1 Original Abstract

*We address the problem of learning good features for understanding video data. We introduce a model that learns latent representations of image sequences from pairs of successive images. The convolutional architecture of our model allows it to scale to realistic image sizes whilst using a compact parametrization. In experiments on the NORB dataset, we show our model extracts latent "flow fields" which correspond to the transformation between the pair of input frames. We also use our model to extract low-level motion features in a multi-stage architecture for action recognition, demonstrating competitive performance on both the KTH and Hollywood2 datasets.*

## 26.2 Main points

# 27 Why does unsupervised pre-training help deep learning? [8]

## 27.1 Original Abstract

*Much recent research has been devoted to learning algorithms for deep architectures such as Deep Belief Networks and stacks of auto-encoder variants, with impressive results obtained in several areas, mostly on vision and language data sets. The best results obtained on supervised learning tasks involve an unsupervised learning component, usually in an unsupervised pre-training phase. Even though these new algorithms have enabled training deep models, many questions remain as to the nature of this difficult learning problem.*

*The main question investigated here is the following: how does unsupervised pre-training work? Answering this questions is important if learning in deep architectures is to be further improved. We propose several explanatory hypotheses and test them through extensive simulations. We empirically show the influence of pre-training with respect to architecture depth, model capacity, and number of training examples. The experiments confirm and clarify the advantage of unsupervised pre-training. The results suggest that unsupervised pre-training guides the learning towards basins of attraction of minima that support better generalization from the training data set; the evidence from these results supports a regularization explanation for the effect of pre-training.*

# 28 Convolutional Deep Belief Networks on CIFAR-10 [25]

## 28.1 Original Abstract

*We describe how to train a two-layer convolutional Deep Belief Network (DBN) on the 1.6 million tiny imagesdataset.When training a convolutional DBN, one must decide what to do with the edge pixels of teh images. Asthe pixels near the edge of an image contribute to the fewest convolutional filter outputs, the model maysee it fit to tailor its few convolutional filters to better model the edge pixels. This is undesirable becaue itusually comes at the expense of a good model for the interior parts of the image. We investigate several waysof dealing with the edge pixels when training a convolutional DBN. Using a combination of locally-connectedconvolutional units and globally-connected units, as well as a few tricks to reduce the effects of overfitting,we achieve state-of-the-art performance in the classification task of the CIFAR-10 subset of the tiny imagesdataset.*

## 28.2 Main points

- Detectors
  - Harris3D
  - Cuboid
  - Hessian

  - Dense sampling

- Descriptors
  - HOG/HOF
  - HOG3D
  - ESURF (extended SURF)

- Datasets
  - KTH actions
    * 6 human action classes
    * walking, jogging, running, boxing, waving and clapping
    * 25 subjects
    * 4 scenarios
    * 2391 video samples
    * $http://www.nada.kth.se/cvap/actions/$
  - UCF sport actions
    * 10 human action classes
    * winging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging, golf swinging and walking
    * 150 video samples
    * $http://crcv.ucf.edu/data/UCF_Sports_Action.php$
  - Hollywood2 actions
    * 12 action classes
    * answering the hone, driving car, eating, fighting, geting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up.
    * 69 Hollywood movies
    * 1707 video samples
    * $http://www.di.ens.fr/\ laptev/actions/hollywood2/$

# 29 Tiled convolutional neural networks. [28]

## 29.1 Original Abstract

*Convolutional neural networks (CNNs) have been successfully applied to many tasks such as digit and object recognition. Using convolutional (tied) weights significantly reduces the number of parameters that have to be learned, and also allows translational invariance to be hard-coded into the architecture. In this paper, we consider the problem of learning invariances, rather than relying on hardcoding. We propose tiled convolution neural networks (Tiled CNNs), which use a regular "tiled " pattern of tied weights that does not require that adjacent hidden units share identical weights, but instead requires only that hidden units k steps away from each other to have tied weights. By pooling over neighboring units, this architecture is able to learn complex invariances (such as scale and rotational invariance) beyond translational invariance. Further, it also enjoys much of CNNs' advantage of having a relatively small number of learned parameters (such as ease of learning and greater scalability). We provide an efficient learning algorithm for Tiled CNNs based on Topographic ICA, and show that learning complex invariant features allows us to achieve highly competitive results for both the NORB and CIFAR-10 datasets.*

## 29.2 Main points

# 30 Building high-level features using large scale unsupervised learning [29]

## 30.1 Original Abstract

*We consider the problem of building high-level, class-specific feature detectors from only unlabeled data. For example, is it possible to learn a face detector using only unlabeled images? To answer this, we train a deep sparse autoencoder on a large dataset of images (the model has 1 billion connections, the dataset has 10 million 200×200 pixel images downloaded from the Internet). We train this network using model parallelism and asynchronous SGD on a cluster with 1,000 machines (16,000 cores) for three days. Contrary to what appears to be a widely-held intuition, our experimental results reveal that it is*

*possible to train a face detector without having to label images as containing a face or not. Control experiments show that this feature detector is robust not only to translation but also to scaling and out-of-plane rotation. We also find that the same network is sensitive to other high-level concepts such as cat faces and human bodies. Starting from these learned features, we trained our network to recognize 22,000 object categories from ImageNet and achieve a leap of 70*

# 31 Stacked convolutional auto-encoders for hierarchical feature extraction [36]

## 31.1 Original Abstract

*We present a novel convolutional auto-encoder (CAE) for unsupervised feature learning. A stack of CAEs forms a convolutional neural network (CNN). Each CAE is trained using conventional on-line gradient descent without additional regularization terms. A max-pooling layer is essential to learn biologically plausible features consistent with those found by previous approaches. Initializing a CNN with filters of a trained CAE stack yields superior performance on a digit (MNIST) and an object recognition (CIFAR10) benchmark.*

# 32 Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis [27]

## 32.1 Original Abstract

*Previous work on action recognition has focused on adapting hand-designed local features, such as SIFT or HOG, from static images to the video domain. In this paper, we propose using unsupervised feature learning as a way to learn features directly from video data. More specifically, we present an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. We discovered that, despite its simplicity, this method performs surprisingly well when combined with deep learning techniques such as stacking and convolution to learn hierarchical rep-*

*resentations. By replacing hand-designed features with our learned features, we achieve classification results superior to all previous published results on the Hollywood2, UCF, KTH and YouTube action recognition datasets. On the challenging Hollywood2 and YouTube action datasets we obtain 53.3*

### 32.2 Main points

## 33 Learning hierarchical features for scene labeling [10]

### 33.1 Original Abstract

*Scene labeling consists of labeling each pixel in an image with the category of the object it belongs to. We propose a method that uses a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel. The method alleviates the need for engineered features, and produces a powerful representation that captures texture, shape, and contextual information. We report results using multiple postprocessing methods to produce the final labeling. Among those, we propose a technique to automatically retrieve, from a pool of segmentation components, an optimal set of components that best explain the scene; these components are arbitrary, for example, they can be taken from a segmentation tree or from any family of oversegmentations. The system yields record accuracies on the SIFT Flow dataset (33 classes) and the Barcelona dataset (170 classes) and near-record accuracy on Stanford background dataset (eight classes), while being an order of magnitude faster than competing approaches, producing a 320×240 image labeling in less than a second, including feature extraction.*

## 34 Gated boltzmann machine in texture modeling [13]

### 34.1 Original Abstract

*In this paper, we consider the problem of modeling complex texture information using undirected probabilistic graphical models. Texture is a special*

*type of data that one can better understand by considering its local structure. For that purpose, we propose a convolutional variant of the Gaussian gated Boltzmann machine (GGBM) [12], inspired by the co-occurrence matrix in traditional texture analysis. We also link the proposed model to a much simpler Gaussian restricted Boltzmann machine where convolutional features are computed as a preprocessing step. The usefulness of the model is illustrated in texture classification and reconstruction experiments.*

### 34.2 Main points

## 35 ImageNet Classification with Deep Convolutional Neural Networks [26]

### 35.1 Original Abstract

*We trained a large, deep convolutional neural network to classify the 1.2 millionhigh-resolution images in the ImageNet LSVRC-2010 contest into the 1000 dif-ferent classes. On the test data, we achieved top-1 and top-5 error rates of 37.5*

### 35.2 Main points

- CNN architecture:
  - 650.000 neurons (60 million parameters)
  - 5 convolutional layers
  - Some of them followed by a max-pooling layer
  - 3 fully-connected layers
  - 1 1000-way softmax
- Dropout regularization method to reduce overfitting in 3 fully-connected layers
- Training time: 5-6 days on two GTX 580 3GB GPUs
- Dataset:
  - ILSVRC-2010

  - Down-sampled images to a fixed resolution of 256x256
  - Substract the mean activity ofver training set from each pixel
- ReLU:
  - $f(x) = \max(0, x)$
  - Faster than tanh
  - ReLU: 6 epochs
  - tanh: 36 more epochs to achieve same performance
- Local Response Normalization
  - 1.2 and 1.4% error reduction
  - Helps generalization
  - $b^i_{x,y} = a^i_{x,y} / \left( k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^\beta$
  - $k = 2, n = 5, \alpha = 10^-4$, and $\beta = 0.75$
- Overlapping Pooling
  - 0.3 and 0.4% error reduction
  - grid $3 \prod 3$
  - stride = 2
  - Overlap each pooling one column pixel
- Overall Architecture
  - 224x224x3 (RGB image)
  - Conv 96 kernels of size 11x11x3 with stride of 4 pixels
  - Response-Normalized and max-pooling
  - Conv 256 kernels of size 5x5x48 with stride of ? pixels
  - Response-Normalized and max-pooling
  - Conv 384 kernels of size 3x3x256
  - Conv 384 kernels of size 3x3x192

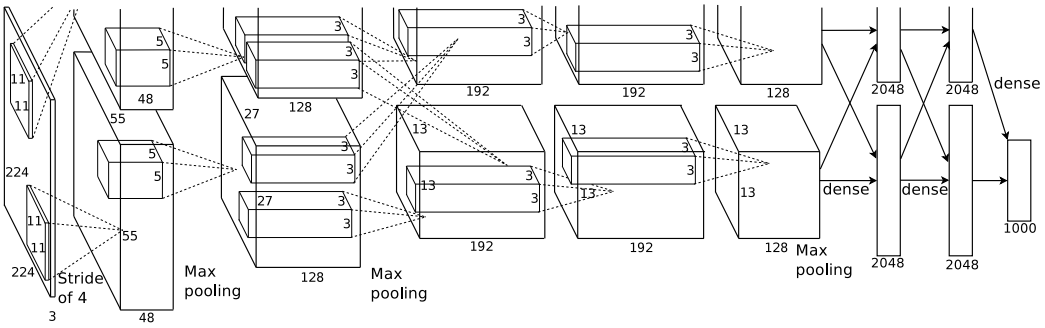


Figure 1: **Architecture of the CNN**

  - Conv 256 kernels of size 3x3x192
  - ¿Response-Normalized? and Max-pooling
  - Fully connected 4096
  - Fully connected 4096
  - Fully connected 1000
  - Softmax

- Data augmentation
  - 0.1 error reduction
  - Original images escaled scaled and croped to 256x256
  - Extract 5 images of 224x224 from corners plus center
  - Mirror horizontally and get 5 more images
  - Augment data altering RGB channels:
    * Perform PCA on RGB throughout the training set
    * Each training image add multiples of PCs with gaussian noise
- Dropout
  - Put to zero the output of neurons with probability 0.5
  - At test time multiply the outputs by 0.5
  - Two first fully-connected layers
  - Solves overfitting

  - Dobules the number of iterations required to ocnverge

- Details of learning
  - batch size = 128
  - momentum 0.9
  - weight decay 0.0005
  - Initial weights from zero-mean Gaussian std=0.01
  - biases = 1 on second, fourth, fifth Conv and fully-connected
  - biases = 0 on the rest

- Evaluation
  - Consider the feature activations induced by an image at the last, 4096-dimensional hidden layer

# 36 The Stanford / Technicolor / Fraunhofer HHI Video [1]

## 36.1 Original Abstract

*Video search has become a very important tool, with the ever-growing size of multimedia collections. This work introduces our Video Semantic Indexing system. Our experiments show that Residual Vectors provide an efficient way of aggregat- ing local descriptors, with complementary gain with respect to BoVW. Also, we show that systems using a limited number of descriptors and machine learning techniques can still be quite effective. Our first participation at the TRECVID evaluation has been very fruitful: our team was ranked 6th in the light version of the Semantic Indexing task.*

## 36.2 Main points

# 37 Improving neural networks by preventing co-adaptation of feature detectors [16]

## 37.1 Original Abstract

*When a large feedforward neural network is trained on a small training set, it typically performs poorly on held-out test data. This "overfitting" is greatly reduced by randomly omitting half of the feature detectors on each training case. This prevents complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors. Instead, each neuron learns to detect a feature that is generally helpful for producing the correct answer given the combinatorially large variety of internal contexts in which it must operate. Random "dropout" gives big improvements on many benchmark tasks and sets new records for speech and object recognition.*

## 37.2 Main points

- Paper about Dropout
- Standard way to reduce test error
  - averaging different models
  - Computationally expensive in training and test
- Dropout
  - Small training set
  - Prevents “overfitting”
  - They use 50%
  - Instead of L2 norm, they set an upper bound for each individual neuron.
  - Mean network : At test time divide all the outgoing weights by 2 to compensate dropout
  - Specific case
    - Single hidden layer network

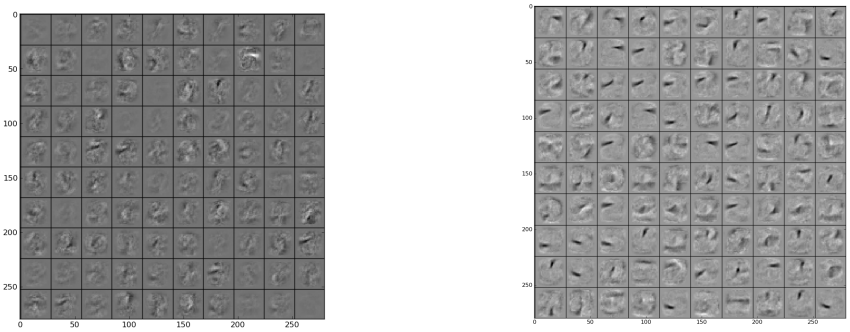

Figure 2: **Visualization of features learned by first layer hidden units** left without dropout and right using dropout

        * N hidden units
        * "Softmax" output
        * 50% dropout
        * during test using mean network
        * Exactly equivalent to taking the geometric mean of the probability distributions over labels predicted by all $2^N$ possible networks

- Results
    - MNIST
        * No dropout : 160 errors
        * Dropbout : 130 errors
        * Dropout + rm random 20% pixels : 110 errors
        * Deep Boltzmann machine : 88 errors
        * + Dropout : 79 errors
    - TIMIT
        * 4 Fully-connected hidden layers 4000 units per layer
        * + 185 "softmax" output units
        * Without dropout : 22.7%
        * Dropout on hidden units : 19.7%

- CIFAR-10
  - Best published : 18.5%
  - 3 Conv+Max-pool 1 Fully : 16.6%
  - + Dropout in last hidden layer : 15.6%
- ImageNet
  - Average of 6 separate models : 47.2%
  - state-of-the-art : 45.7%
  - 5 Conv+Max-pool
  - + 2 Fully
  - + 1000 "softmax"
  - Without dropout : 48.6%
  - Dropout in the 6th : 42.4%
- Reuters
  - 2 fully of 2000 hidden units
  - Without dropout : 31.05%
  - Dropout : 29.62%

# 38 Recognizing 50 human action categories of web videos [41]

## 38.1 Original Abstract

*Action recognition on large categories of unconstrained videos taken from the web is a very challenging problem compared to datasets like KTH (6 actions), IXMAS (13 actions), and Weizmann (10 actions). Challenges like camera motion, different viewpoints, large interclass variations, cluttered background, occlusions, bad illumination conditions, and poor quality of web videos cause the majority of the state-of-the-art action recognition approaches to fail. Also, an increased number of categories and the inclusion of actions with high confusion add to the challenges. In this paper, we propose using the scene context information obtained from moving and stationary pixels in the key frames, in conjunction with motion features, to solve the action recognition problem on a large (50 actions) dataset with videos from the web. We perform a combination of early and late fusion on multiple features to handle the very large*

*number of categories. We demonstrate that scene context is a very important feature to perform action recognition on very large datasets. The proposed method does not require any kind of video stabilization, person detection, or tracking and pruning of features. Our approach gives good performance on a large number of action categories; it has been tested on the UCF50 dataset with 50 action categories, which is an extension of the UCF YouTube Action (UCF11) dataset containing 11 action categories. We also tested our approach on the KTH and HMDB51 datasets for comparison.*

# 39 Mitosis detection in breast cancer histology images with deep neural networks [6]

## 39.1 Original Abstract

*We use deep max-pooling convolutional neural networks to detect mitosis in breast histology images. The networks are trained to classify each pixel in the images, using as context a patch centered on the pixel. Simple postprocessing is then applied to the network output. Our approach won the ICPR 2012 mitosis detection competition, outperforming other contestants by a significant margin.*

# 40 Understanding Deep Architectures using a Recursive Convolutional Network [7]

## 40.1 Original Abstract

*A key challenge in designing convolutional network models is sizing them appropriately. Many factors are involved in these decisions, including number of layers, feature maps, kernel sizes, etc. Complicating this further is the fact that each of these influence not only the numbers and dimensions of the activation units, but also the total number of parameters. In this paper we focus on assessing the independent contributions of three of these linked variables: The numbers of layers, feature maps, and parameters. To accomplish this, we employ a recursive convolutional network whose weights are tied between layers; this allows us to vary each of the three factors in a controlled setting. We find that while increasing the numbers of layers and parameters each have*

*clear benefit, the number of feature maps (and hence dimensionality of the representation) appears ancillary, and finds most of its benefit through the introduction of more weights. Our results (i) empirically confirm the notion that adding layers alone increases computational power, within the context of convolutional layers, and (ii) suggest that precise sizing of convolutional feature map dimensions is itself of little concern; more attention should be paid to the number of parameters in these layers instead.*

### 40.2 Main points

- Deeper models are preferred over shallow ones
- Performance is independent of the number of units, when depth and parameters remains constant
- Recurrent Neural Network:
  - Convolutional architecture
  - all layers same number of feature maps
  - weights are tied across layers
  - ReLU in all layers
  - Max-pooling with non-overlaping windows

## 41 Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps [46]

### 41.1 Original Abstract

*This paper addresses the visualisation of image classification models, learnt using deep Convolutional Networks (ConvNets). We consider two visualisation techniques, based on computing the gradient of the class score with respect to the input image. The first one generates an image, which maximises the class score [Erhan et al., 2009], thus visualising the notion of the class, captured by a ConvNet. The second technique computes a class saliency map, specific to a given image and class. We show that such maps*

*can be employed for weakly supervised object segmentation using classification ConvNets. Finally, we establish the connection between the gradient-based ConvNet visualisation methods and deconvolutional networks [Zeiler et al., 2013].*

### 41.2 Main points

## 42 Visualizing and Understanding Convolutional Networks [52]

### 42.1 Original Abstract

*Large Convolutional Network models have recently demonstrated impressive classification performance on the ImageNet benchmark. However there is no clear understanding of why they perform so well, or how they might be improved. In this paper we address both issues. We introduce a novel visualization technique that gives insight into the function of intermediate feature layers and the operation of the classifier. We also perform an ablation study to discover the performance contribution from different model layers. This enables us to find model architectures that outperform Krizhevsky et.al. on the ImageNet classification benchmark. We show our ImageNet model generalizes well to other datasets: when the softmax classifier is retrained, it convincingly beats the current state-of-the-art results on Caltech-101 and Caltech-256 datasets.*

## 43 Action and event recognition with Fisher vectors on a compact feature set [39]

### 43.1 Original Abstract

*Action recognition in uncontrolled video is an important and challenging computer vision problem. Recent progress in this area is due to new local features and models that capture spatio-temporal structure between local features, or human-object interactions. Instead of working towards more complex models, we focus on the low-level features and their encoding. We evaluate the use of Fisher vectors as an alternative to bag-of-word histograms to aggregate a*

*small set of state-of-the-art low-level descriptors, in combination with linear classifiers. We present a large and varied set of evaluations, considering (i) classification of short actions in five datasets, (ii) localization of such actions in feature-length movies, and (iii) large-scale recognition of complex events. We find that for basic action recognition and localization MBH features alone are enough for state-of-the-art performance. For complex events we find that SIFT and MFCC features provide complementary cues. On all three problems we obtain state-of-the-art results, while using fewer features and less complex models.*

### 43.2 Main points

## 44 TRECVID 2013 – An Introduction to the Goals , Tasks , Data , Evaluation Mechanisms , and Metrics [40]

### 44.1 Original Abstract

*None*

### 44.2 Main points

## 45 Quaero at TRECVid 2013 : Semantic Indexing [42]

### 45.1 Original Abstract

*The Quaero group is a consortium of French and German organizations working on Multimedia Indexing and Retrieval. LIG, INRIA and KIT participated to the semantic indexing task and LIG participated to the organization of this task. This paper describes these participations. For the semantic indexing task, our approach uses a six-stages processing pipelines for computing scores for the likelihood of a video shot to contain a target concept. These scores are then used for producing a ranked list of images or shots that are the most likely to contain the target concept. The pipeline is composed of the following steps: descriptor extraction, descriptor optimization, classi cation,*

*fusion of descriptor variants, higher-level fusion, and re-ranking. We used a number of di erent descriptors and a hierarchical fusion strategy. We also used conceptual feedback by adding a vector of classi cation score to the pool of descriptors. The best Quaero run has a Mean Inferred Average Precision of 0.2692, which ranked us 3rd out of 16 participants. We also organized the TRECVid SIN 2012 collaborative annotation.*

### 45.2 Main points

## 46 MediaMill at TRECVID 2013: Searching Concepts, Objects, Instances and Events in Video [47]

### 46.1 Original Abstract

*In this paper we summarize our TRECVID 2013 video retrieval experiments. The MediaMill team participated in four tasks: concept detection, object localization, in- stance search, and event recognition. For all tasks the starting point is our top-performing bag-of-words system of TRECVID 2008-2012, which uses color SIFT descrip- tors, average and difference coded into codebooks with spa- tial pyramids and kernel-based machine learning. New this year are concept detection with deep learning, concept detec- tion without annotations, object localization using selective search, instance search by reranking, and event recognition based on concept vocabularies. Our experiments focus on es- tablishing the video retrieval value of the innovations. The 2013 edition of the TRECVID benchmark has again been a fruitful participation for the MediaMill team, resulting in the best result for concept detection, concept detection with- out annotation, object localization, concept pair detection, and visual event recognition with few examples.*

### 46.2 Main points

## 47 Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video [51]

### 47.1 Original Abstract

*We propose a novel approach to boost the performance of generic object detectors on videos by learning video-specific features using a deep neural network. The insight behind our proposed approach is that an object appearing in different frames of a video clip should share similar features, which can be learned to build better detectors. Unlike many supervised detector adaptation or detection-by-tracking methods, our method does not require any extra annotations or utilize temporal correspondence. We start with the high-confidence detections from a generic detector, then iteratively learn new video-specific features and refine the detection scores. In order to learn discriminative and compact features, we propose a new feature learning method using a deep neural network based on auto en-coders. It differs from the existing unsupervised feature learning methods in two ways: first it optimizes both discriminative and generative properties of the features simultaneously, which gives our features better discriminative ability, second, our learned features are more compact, while the unsupervised feature learning methods usually learn a redundant set of over-complete features. Extensive experimental results on person and horse detection show that significant performance improvement can be achieved with our proposed method.*

## 48 Learned versus Hand-Designed Feature Representations for 3d Agglomeration [3]

### 48.1 Original Abstract

*For image recognition and labeling tasks, recent results suggest that machine learning methods that rely on manually specified feature representations may be outperformed by methods that automatically derive feature representations based on the data. Yet for problems that involve analysis of 3d objects, such as mesh segmentation, shape retrieval, or neuron fragment agglomeration, there*

*remains a strong reliance on hand-designed feature descriptors. In this paper, we evaluate a large set of hand-designed 3d feature descriptors alongside features learned from the raw data using both end-to-end and unsupervised learning techniques, in the context of agglomeration of 3d neuron fragments. By combining unsupervised learning techniques with a novel dynamic pooling scheme, we show how pure learning-based methods are for the first time competitive with hand-designed 3d shape descriptors. We investigate data augmentation strategies for dramatically increasing the size of the training set, and show how combining both learned and hand-designed features leads to the highest accuracy.*

# 49 OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks [43]

## 49.1 Original Abstract

*We present an integrated framework for using Convolutional Networks for classification, localization and detection. We show how a multiscale and sliding window approach can be efficiently implemented within a ConvNet. We also introduce a novel deep learning approach to localization by learning to predict object boundaries. Bounding boxes are then accumulated rather than suppressed in order to increase detection confidence. We show that different tasks can be learned simultaneously using a single shared network. This integrated framework is the winner of the localization task of the ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013) and obtained very competitive results for the detection and classifications tasks. In post-competition work, we establish a new state of the art for the detection task. Finally, we release a feature extractor from our best model called OverFeat.*

## 49.2 Main points

- Framework for using CNN
  - classification
  - localization

  - detection

- Winner on localization task of ILSVRC2013
- ConvNets are trained enterily with the raw pixels
- Other approaches for detection and localization
- appling a sliding window over multiples scales
- ...
- ...

# 50 Learning Deep Face Representation [9]

## 50.1 Original Abstract

*Face representation is a crucial step of face recognition systems. An optimal face representation should be discriminative, robust, compact, and very easy-to-implement. While numerous hand-crafted and learning-based representations have been proposed, considerable room for improvement is still present. In this paper, we present a very easy-to-implement deep learning framework for face representation. Our method bases on a new structure of deep network (called Pyramid CNN). The proposed Pyramid CNN adopts a greedy-filter-and-down-sample operation, which enables the training procedure to be very fast and computation-efficient. In addition, the structure of Pyramid CNN can naturally incorporate feature sharing across multi-scale face representations, increasing the discriminative ability of resulting representation. Our basic network is capable of achieving high recognition accuracy (85.8*

## 50.2 Main points

- New deep structure Pyramid CNN
- Labeled Faces in the Wild (LFW)
  - $>$ 13.000 faces
  - 1680 of the people have two or more distinct photos

  - Detected by Viola-Jones detector
  - http://vis-www.cs.umass.edu/lfw/
- State-of-the-art performance on LFW benchmark (97.3%)
- Good face representation
  - Identity-preserving: Same person pictures close in feature space
  - Abstract and Compact: from high to low dimensionality
  - Uniform and Automatic: NO hand-crafted and hard-wired parts
- Pyramid CNN
  - ID-preserving Representation Learning: Loss functions measures distance in output feature space
  - Convolutions and Down-sampling
  - Deeper give best results, but increases rapidly the training time
  - Each CNN own private output layer and gets the input from the previous shared layer
  - Only the output of the last level network is used for the represetnation
  - The rest of the outputs is just for training
- Results
  - 164 incorrect predictions
  - Some of them are incorrectly labeled
  - Others are very difficult for humans, because of the age or pose
  - On LFW benchmark achieves state-of-the-art and close to human on croped images
- With ROC curve as a mesure there is an improvement of 0.07-0.12 with Baseline
- Face recognition does not contemplate affine transformations or perspectives,
- Can be difficult to apply in task such as ImageNet, where the object can be in any place and position

# 51 Towards Real-Time Image Understanding with Convolutional Networks [11]

## 51.1 Original Abstract

*One of the open questions of artificial computer vision is how to produce good internal representations of the visual world. What sort of internal representation would allow an artificial vision system to detect and classify objects into categories, independently of pose, scale, illumination, conforma- tion, and clutter? More interestingly, how could an artificial vision system learn appropriate internal representations automatically, the way animals and humans seem to learn by simply looking at the world? Another related question is that of computational tractability, and more precisely that of computational efficiency. Given a good visual represen- tation, how efficiently can it be trained, and used to encode new sensorial data. Efficiency has several dimensions: power requirements, processing speed, and memory usage. In this thesis I present three new contributions to the field of computer vision: (1) a multiscale deep convolutional network architecture to easily capture long-distance relationships between input variables in image data, (2) a tree-based algorithm to efficiently explore multiple segmentation can- didates, to produce maximally confident semantic segmentations of images, (3) a custom dataflow computer architecture optimized for the computation of convolutional networks, and similarly dense image processing models. All three contributions were produced with the common goal of getting us closer to real-time image understanding. Scene parsing consists in labeling each pixel in an image with the category of the object it belongs to. In the first part of this thesis, I propose a method that uses a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel. The method alleviates the need for engineered features. Inparallel to feature extraction, a tree of segments is computed from a graph of pixel dissimilarities. The feature vectors associated with the segments covered by each node in the tree are aggregated and fed to a classifier which produces an estimate of the distribution of object categories contained in the segment. A subset of tree nodes that cover the image are then selected so as to maximize the average "purity" of the class distributions, hence maximizing the overall likelihood that each segment contains a single object. The system yields record accuracies on several public*

*benchmarks. The computation of convolutional networks, and related models heavily relies on a set of basic operators that are particularly fit for dedicated hardware implementations. In the second part of this thesis I introduce a scalable dataflow hardware architecture optimized for the computation of general-purpose vision algorithms—neuFlow —and a dataflow compiler—luaFlow —that transforms high-level flow-graph representations of these algorithms into machine code for neuFlow. This system was designed with the goal of providing real-time detection, categorization and localization of objects in complex scenes, while consuming 10 Watts when implemented on a Xilinx Virtex 6 FPGA platform, or about ten times less than a lap- top computer, and producing speedups of up to 100 times in real-world applications (results from 2011).*

# 52 Spectral Networks and Deep Locally Connected Networks on Graphs [4]

## 52.1 Original Abstract

*Convolutional Neural Networks are extremely efficient architectures in image and audio recognition tasks, thanks to their ability to exploit the local translational invariance of signal classes over their domain. In this paper we consider possi- ble generalizations of CNNs to signals defined on more general domains without the action of a translation group. In particular, we propose two constructions, one based upon a hierarchical clustering of the domain, and another based on the spectrum of the graph Laplacian. We show through experiments that for low- dimensional graphs it is possible to learn convolutional layers with a number of parameters independent of the input size, resulting in efficient deep architectures.*

# 53 Large-scale Video Classification with Convolutional Neural Networks [22]

## 53.1 Original Abstract

*Convolutional Neural Networks (CNNs) have been es-tablished as a powerful class of models for image recog-nition problems. Encouraged by these results,*

*we pro-vide an extensive empirical evaluation of CNNs on large- scale video classification using a new dataset of 1 millionYouTube videos belonging to 487 classes. We study mul-tiple approaches for extending the connectivity of a CNNin time domain to take advantage of local spatio-temporalinformation and suggest a multiresolution, foveated archi-tecture as a promising way of speeding up the training.Our best spatio-temporal networks display significant per-formance improvements compared to strong feature-basedbaselines (55.3*

## 53.2 Main points

- Compare different CNN architectures for video classification
- Create a new dataset with 1 million of YouTube sport videos and 487 classes
- They required one month of training
- Multiresolution CNNs: New CNN with low resolution context and high resolution center
  - Context stream: seems to learn color filters
  - Fovea stream: learns grayscale features
- Compare with and without pretraining on other dataset UCF-101
- Architectures (increasing spatio-temporal relations)
  - Single frame: Classify with one single shot
  - Late Fusion: Classify with separate-in-time shots
  - Early Fusion: Classify with adjacent shots merging on first convolution layer
  - Slow Fusion: Classify with adjacent shots progressively mergin in upper layers
- Results (best models):
  - clip Hit, Video Hit, Video Hit top5
  - 42.4 60.0 78.5 Single-Frame + Multiresolution
  - 41.9 60.9 80.2 Slow Fusion

- Results on UCF-101 with pretraining:
  - 41.3 No pretraining
  - 64.1 Fine-tune top layer
  - 65.4 Fine-tune top 3 layers
  - 62.2 Fine-tune all layers
- Conclusions:
  - From video classification can be derived that camera movements deteriorate the predictions
  - Single frame gives very good results
- Further work:
  - Apply some filter for camera movements
  - Explore RNN from clip-level into video-level

# References


[1] A F De Araujo, F Silveira, H Lakshman, J Zepeda, A Sheth, and B Girod. The Stanford / Technicolor / Fraunhofer HHI Video. 2012.

[2] Yoshua Bengio. *Learning Deep Architectures for AI*, volume 2. 2009.

[3] JA Bogovic, GB Huang, and Viren Jain. Learned versus Hand-Designed Feature Representations for 3d Agglomeration. *arXiv preprint arXiv:1312.6159*, pages 1–14, 2013.

[4] Joan Bruna, Arthur Szlam, Wojciech Zaremba, and Yann LeCun. Spectral Networks and Deep Locally Connected Networks on Graphs. pages 1–14, 2014.

[5] P Le Callet. A convolutional neural network approach for objective video quality assessment. *Neural Networks, IEEE ...*, 5:1316–1327, 2006.

[6] DC Cireşan and Alessandro Giusti. Mitosis detection in breast cancer histology images with deep neural networks. *Medical Image ...*, 2013.

[7] David Eigen, Jason Rolfe, Rob Fergus, and Y LeCun. Understanding Deep Architectures using a Recursive Convolutional Network. *arXiv preprint arXiv:1312.1847*, pages 1–9, 2013.

[8] D Erhan, Yoshua Bengio, and Aaron Courville. Why does unsupervised pre-training help deep learning? *... of Machine Learning ...*, 9(2007):201–208, 2010.

[9] Haoqiang Fan, Zhimin Cao, Yunin Jiang, Qi Yin, C Doudou, and Chinchilla Doudou. Learning Deep Face Representation. *arXiv preprint arXiv:1403.2802*, pages 1–10, 2014.

[10] C Farabet, Camille Couprie, Laurent Najman, and Y LeCun. Learning hierarchical features for scene labeling. 8:1915–1929, 2012.

[11] Clément Farabet. *Towards Real-Time Image Understanding with Convolutional Networks*. PhD thesis, Université Paris-Est, 2014.

[12] Kunihiko Fukushima. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. 202, 1980.

[13] Tele Hao, Tapani Raiko, Alexander Ilin, and Juha Karhunen. Gated boltzmann machine in texture modeling. *... Neural Networks and Machine ...*, 2012.

[14] GE Hinton, Simon Osindero, and YW Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 2006.

[15] GE Hinton and RR Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(July):504–507, 2006.

[16] GE Hinton, N Srivastava, Alex Krizhevsky, I Sutskever, and RR Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv: ...*, pages 1–18, 2012.

[17] Geoffrey Hinton. To recognize shapes, first learn to generate images. *Progress in brain research*, 2007.

[18] P O Hoyer and a Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network (Bristol, England)*, 11(3):191–210, August 2000.

[19] DH Hubel and TN Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, pages 106–154, 1962.

[20] DH Hubel and TN Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, pages 215–243, 1968.

[21] A Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 1720:1705–1720, 2000.

[22] Andrej Karpathy, G Toderici, S Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. *vision.stanford.edu*, 2014.

[23] Koray Kavukcuoglu, Pierre Sermanet, Y-lan Boureau, Yann LeCun, Karol Gregor, and Michaël Mathieu. Learning Convolutional Feature Hierarchies for Visual Recognition. *NIPS*, (1):1–9, 2010.

[24] JC King. Why color management? *9th Congress of the International Color ...*, 2002.

[25] Alex Krizhevsky. Convolutional Deep Belief Networks on CIFAR-10. pages 1–9, 2010.

[26] Alex Krizhevsky, I Sutskever, and GE Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, pages 1–9, 2012.

[27] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *Cvpr 2011*, pages 3361–3368, June 2011.

[28] QV Le, Jiquan Ngiam, Zhenghao Chen, DJ hao Chia, and PW Koh. Tiled convolutional neural networks. *NIPS*, pages 1–9, 2010.

[29] QV Le, MA Ranzato, R Monga, and Matthieu Devin. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv: ...*, 2011.

[30] Y LeCun. Generalization and network design strategies. *Connections in Perspective. North-Holland, ...*, 1989.

[31] Y LeCun and Y Bengio. Convolutional networks for images, speech, and time series. *... handbook of brain theory and neural networks*, pages 1–14, 1995.

[32] Y LeCun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel. Backpropagation applied to handwritten zip code recognition. *Neural ...*, 1989.

[33] Y LeCun and L Bottou. Gradient-based learning applied to document recognition. *Proceedings of the ...*, 1998.

[34] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, 2009.

[35] M Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. *Computer Vision and ...*, (i):2929–2936, 2009.

[36] Jonathan Masci, Ueli Meier, D Cireşan, and J Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks ...*, pages 52–59, 2011.

[37] Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pan Wei Koh, and Andrew Y. Ng. Tiled convolutional neural networks. *Advances in Neural ...*, pages 1–9, 2010.

[38] Mohammad Norouzi, Mani Ranjbar, and Greg Mori. Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning. *Computer Vision and Pattern ...*, pages 2735–2742, 2009.

[39] D Oneata, Jakob Verbeek, and C Schmid. Action and event recognition with Fisher vectors on a compact feature set. *... Conference on Computer ...*, 2013.

[40] Paul Over, George Awad, Jon Fiscus, and Greg Sanders. TRECVID 2013 – An Introduction to the Goals , Tasks , Data , Evaluation Mechanisms , and Metrics. 2013.

[41] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, November 2012.

[42] Bahjat Safadi, Nadia Derbas, Abdelkader Hamadi, Thi-thu-thuy Vuong, Han Dong, Philippe Mulhem, and Georges Qu. Quaero at TRECVid 2013 : Semantic Indexing. 2013.

[43] Pierre Sermanet, David Eigen, X Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv preprint arXiv: ...*, pages 1–16, 2014.

[44] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–26, March 2007.

[45] P Simard, Dave Steinkraus, and JC Platt. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. *ICDAR*, 2003.

[46] Karen Simonyan, A Vedaldi, and A Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*, pages 1–8, 2013.

[47] CGM Snoek and KEA van de Sande. MediaMill at TRECVID 2013: Searching Concepts, Objects, Instances and Events in Video. *... of TRECVID*, 2013.

[48] GW Taylor, Rob Fergus, Y LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. *Computer Vision–ECCV 2010*, 2010.

[49] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for

action recognition. *Procedings of the British Machine Vision Conference 2009*, pages 124.1–124.11, 2009.

[50] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1168–1175, 2008.

[51] Yang Yang, Guang Shu, and Mubarak Shah. Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1650–1657, June 2013.

[52] MD Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *arXiv preprint arXiv:1311.2901*, 2013.