Visual Feature Learning and Representation

Qingshan Liu

Nanjing University of Information Science & Technology

11. 5. 2016

What can we read from this story?

What Can We Read From Face Images?

Visual Recognition = Feature + Classifier

Global feature vs. Local feature ?

From Shiguang Shan

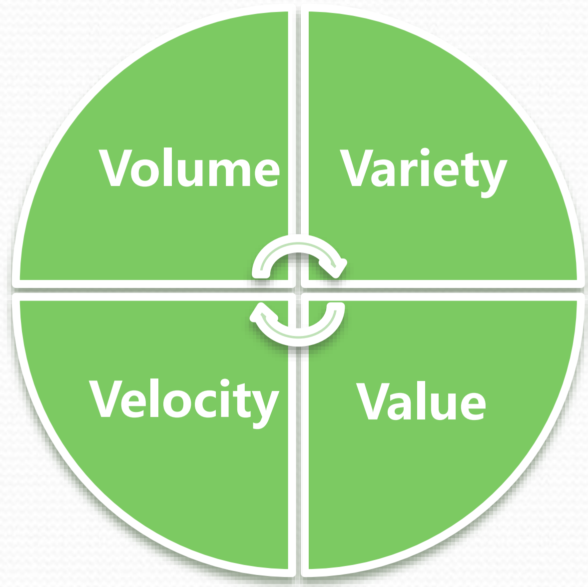
Challenges

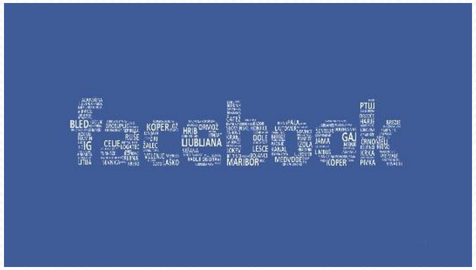
- There have 100 million survellience cameras distributed in the word, which will produce **2.3 ZB** (10^{21}) video data

- Youtube will increase over **72 hours** video data in each minute

- Face book has over **300 billion** images

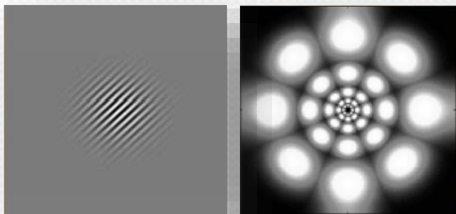
- ……



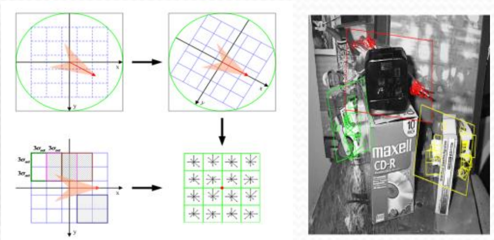


Visual Feature Representation

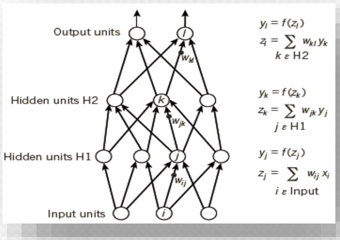
Gabor Filter Bank



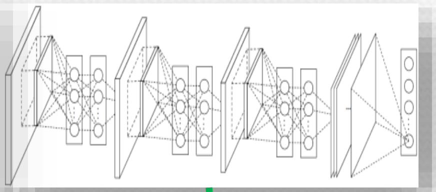
SIFT
(Scale-invariant feature transform)



DBN

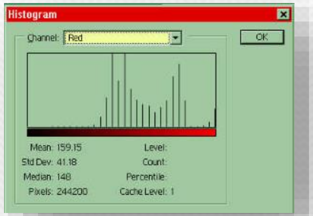


Nested Network

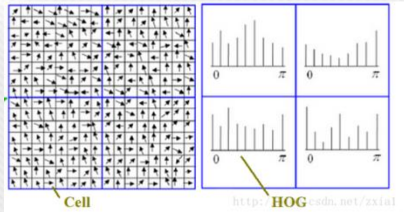


1990s | 2000s | 2004 | 2005 | 2006 | 2010~

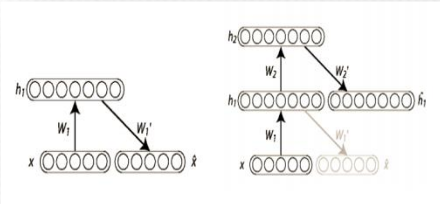
color histogram



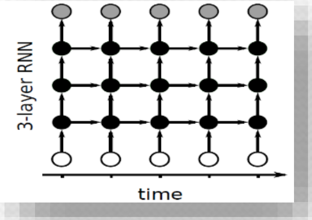
HOG



Stacked Encoder



Recurrent Network



Low level feature | Hand-crafted feature | Deep feature

Data driven

Complicated

High dimension issue

2000

2005~2013

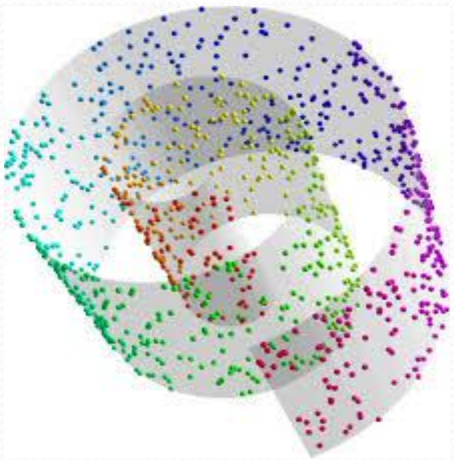
2013~

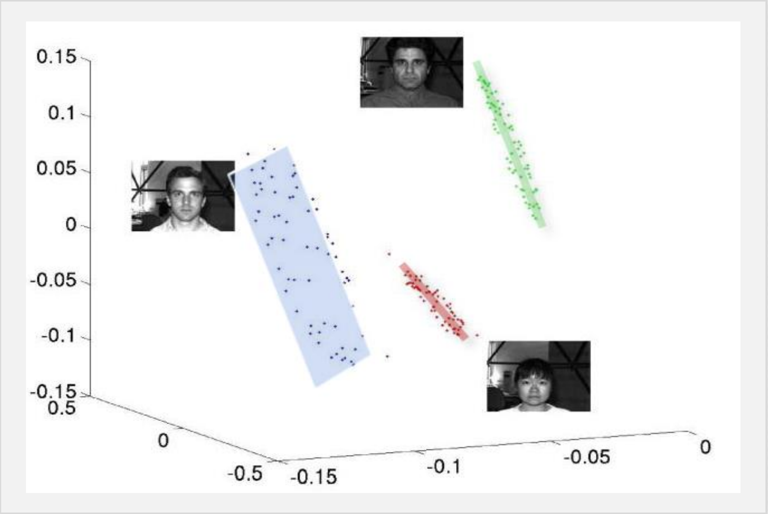
Hundreds of thousands pixels | Million pixels | ten million pixels

Sensor driven

High dimension

David L. Donoho, High-dimensional data analysis: The curses and blessings of dimensionality. ***Aide-Memoire of a Lecture*** **at (2000)**





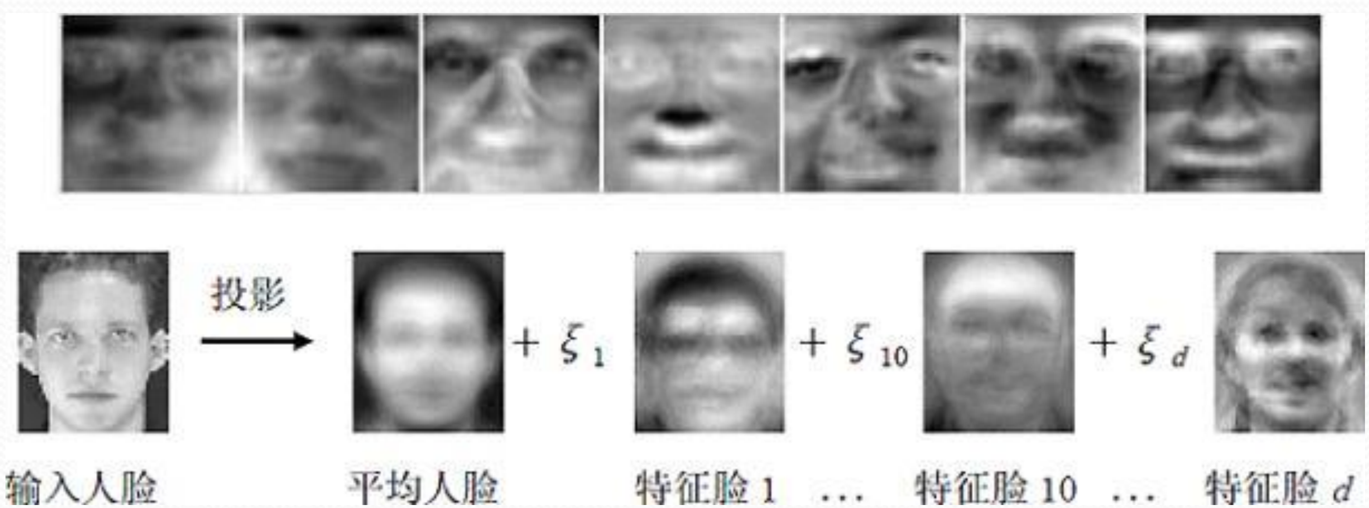
How to learn the low dimensional feature representation

Subspace Learning

- **Learn a low dimensional subspace projection to handle the high-dimensional data**

$$y = A^T x$$

$$x \in R^D, \quad A \in R^{D\times d}, \quad y \in d, \quad d < D.$$



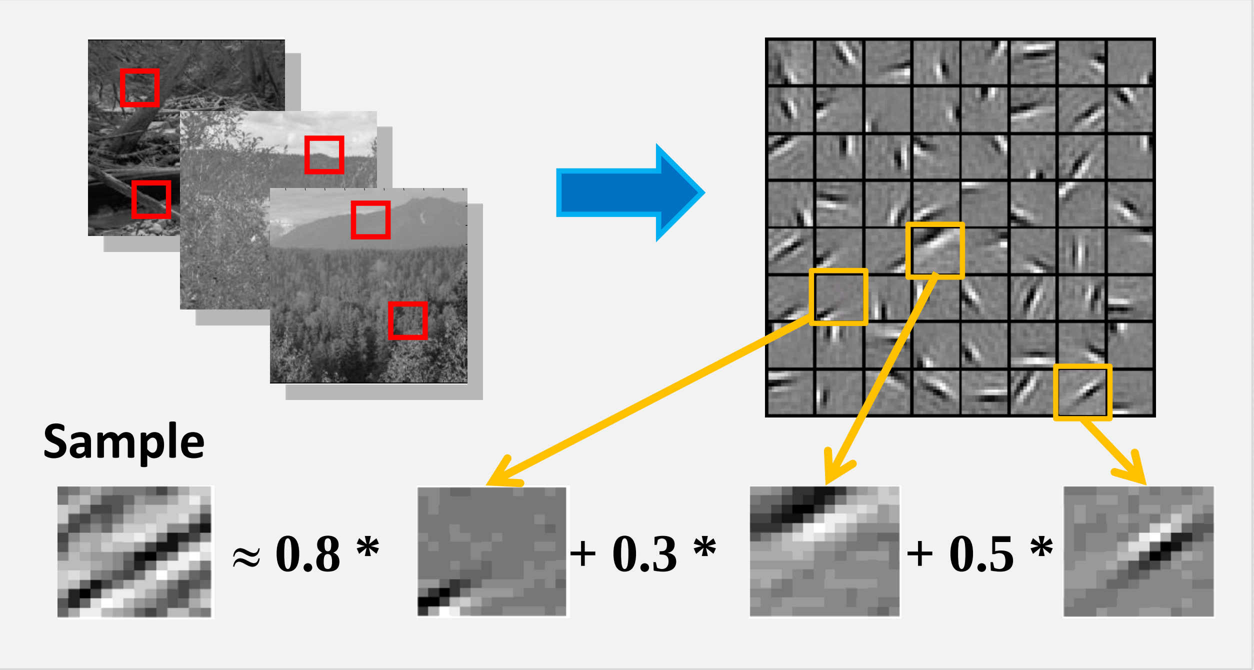
Subspace Learning

- **Linear subspace**: A is a linear transformation
  for example: PCA, LDA,…
- **Kernel based nonlinear subspace**: combining the nonlinear kernel trick with linear subspace
  for example: KPCA, KLDA,…
- **Manifold subspace**
  for example: LLE, ISOMap,…

Sparse feature representation

Simple

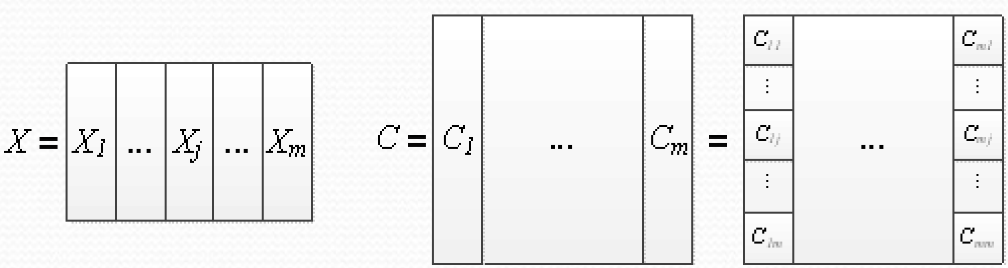
Reliable

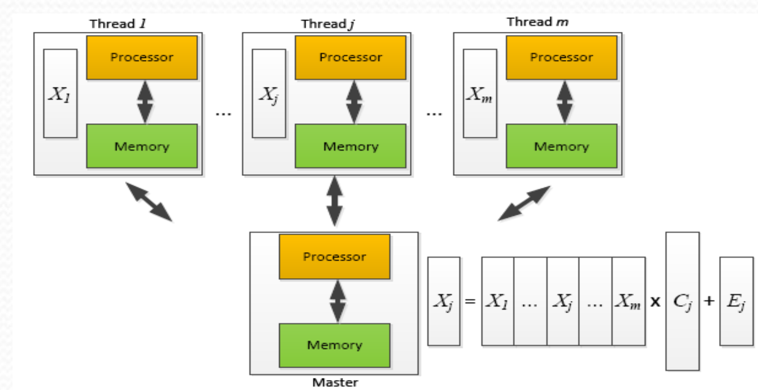


$$\min_{A,X} \|X - AZ\|_2^2 + \lambda \|Z\|_1$$

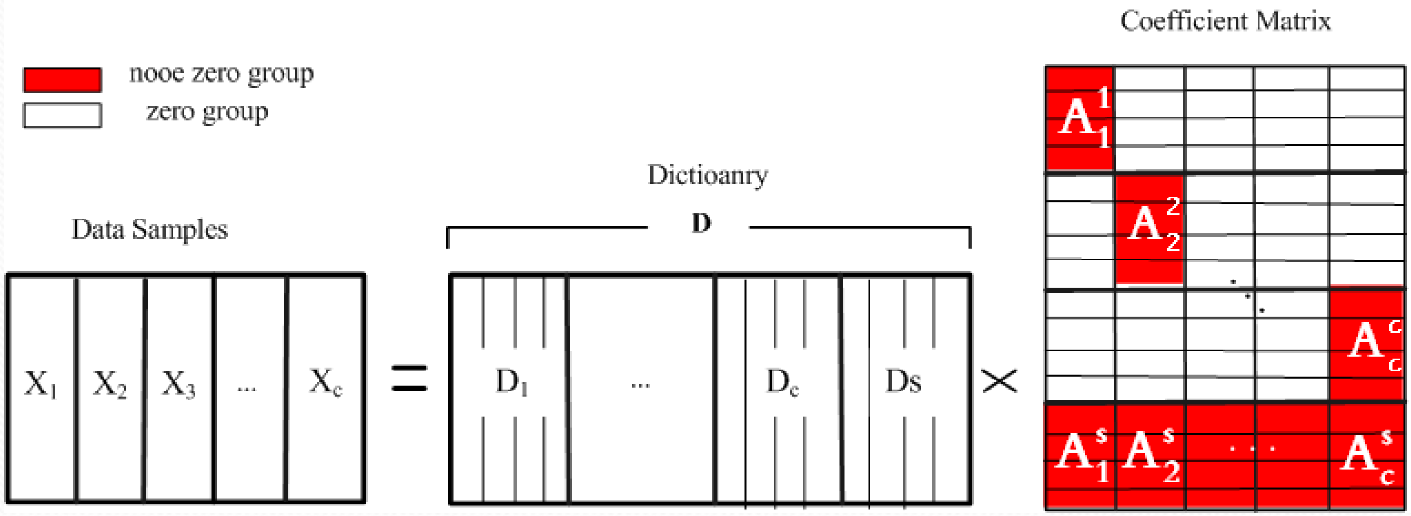
Sparse Representation

- **Learning Discriminative Dictionary for Group Sparse Representation (IEEE T-IP 2014)**
- **Newton Greedy Pursuit: a Quadratic Approximation Method for Sparsity-Constrained Optimization, (CVPR 2014).**
- **Decentralized Robust Subspace Clustering (AAAI 2016)**
- **Efficient k-Support-Norm Regularized Minimization via Fully Corrective Frank-Wolfe Method (IJCAI 2016)**
- **Efficient λ^2 Kernel Linearization via Random Feature Maps (IEEE T-NNLS 2016)**
- **Blessing of Dimensionality: Recovering Mixture Data via Dictionary Pursuit, (IEEE T-PAMI 2016)**





Learning Discriminative Dictionary for Group Sparse Representation (IEEE T-IP 2014)



$$\min_{\mathbf{D},\mathbf{A}} \left\{ \begin{array}{l} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 + \sum_{i=1}^{c} \left\|\mathbf{X}_i - \mathbf{D}_i\mathbf{A}_i^i - \mathbf{D}_s\mathbf{A}_i^s\right\|_F^2 \\ +\eta \sum_{i \neq j} \left\|\mathbf{D}_i^T\mathbf{D}_j\right\|_F^2 + \lambda(\|\mathbf{A}\|_{\omega,2,1}) \end{array} \right\}$$
$$s.t. \ \|\mathbf{d}_i\|_2^2 \leq 1, \quad i = 1, \ldots, m.$$

TABLE I

RECOGNITION RATES OF VARIOUS METHODS ON THE AR FACE DATABASE

Algorithms	SRC	D-KSVD	DLSI	MFL	FDDL
Recognition Rate	0.889 $\pm$0.015	0.856 $\pm$0.012	0.896 $\pm$0.009	0.9126 $\pm$0.011	0.920 $\pm$0.009
	LC-KSVD	DL-COPAR	DGSDL$^{\#}$	DGSDL	
Recognition Rate	0.9396 $\pm$0.008	0.9401 $\pm$0.007	0.9364 $\pm$0.006	**0.9442 $\pm$0.005**	

TABLE II

RECOGNITION RATES OF VARIOUS METHODS ON THE CMU PIE DATABASE

Algorithms	SRC	D-KSVD	DLSI	MFL	FDDL
Recognition Rate	0.956 $\pm$0.009	0.937 $\pm$0.01	0.942 $\pm$0.01	0.953 $\pm$0.009	0.967 $\pm$0.008
	LC-KSVD	DL-COPAR	DGSDL$^{\#}$	DGSDL	
Recognition Rate	0.972 $\pm$0.006	0.9723 $\pm$0.007	0.9635 $\pm$0.005	**0.9726 $\pm$0.004**	

TABLE III

RECOGNITION RATES OF VARIOUS METHODS ON THE EXTENDED YALE B

Algorithms	SRC	D-KSVD	DLSI	MFL	FDDL
Recognition Rate	0.902 $\pm$0.012	0.756 $\pm$0.017	0.890 $\pm$0.014	0.9371 $\pm$0.011	0.9391 $\pm$0.01
	LC-KSVD	DL-COPAR	DGSDL$^{\#}$	DGSDL	
Recognition Rate	0.947 $\pm$0.009	0.951 $\pm$0.008	0.9371 $\pm$0.009	**0.9572 $\pm$0.007**	

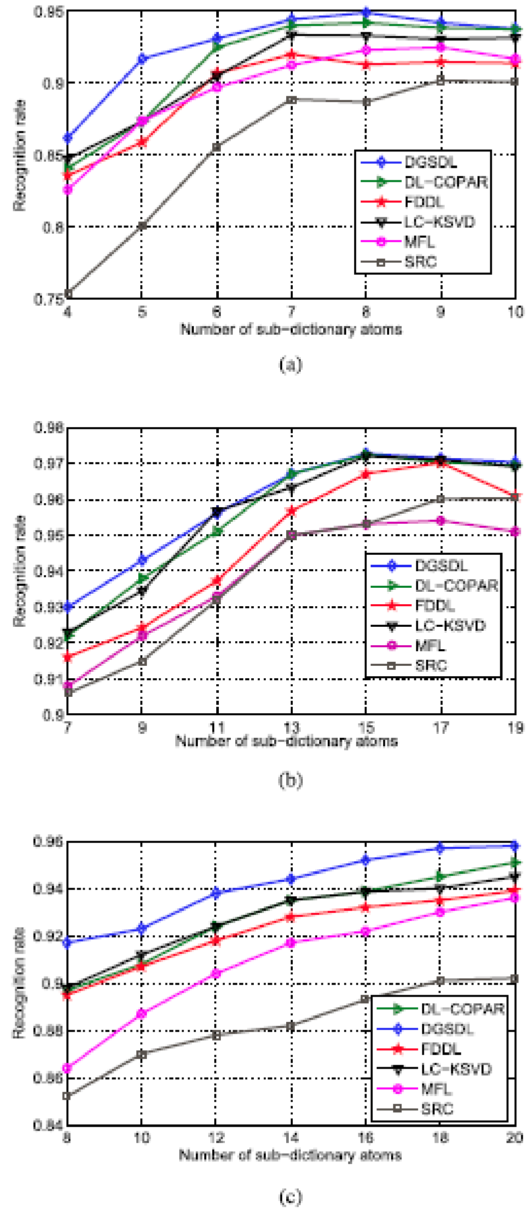
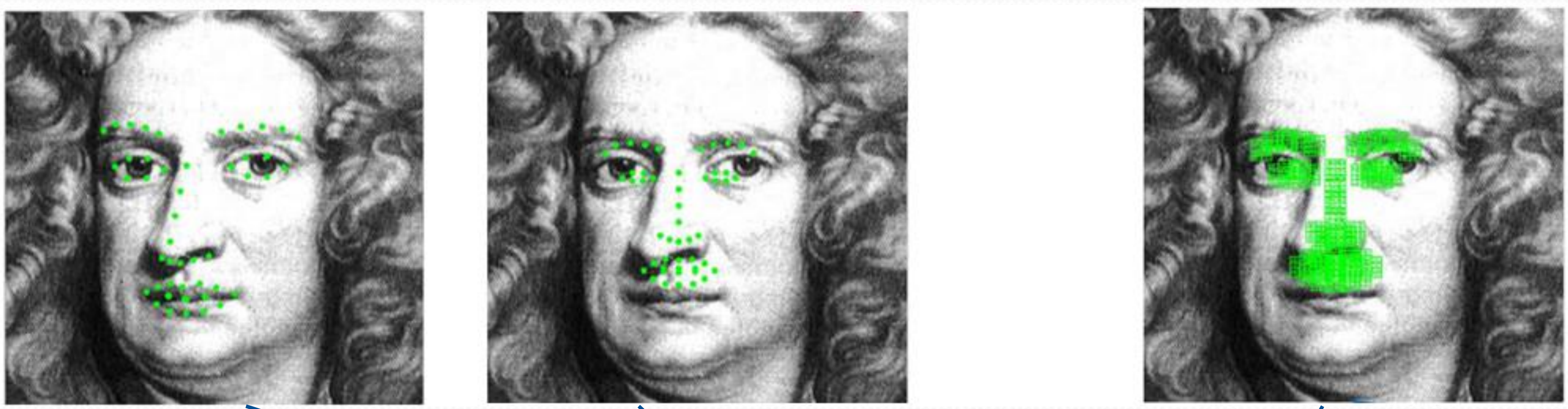


Fig. 8. Recognition rate plots of SRC, MFL, DGSDL, LC-KSVD, FDDL and DL-COPAR versus different number of sub-dictionary atoms. (a) AR. (b) CMU PIE. (c) Extended Yale B.

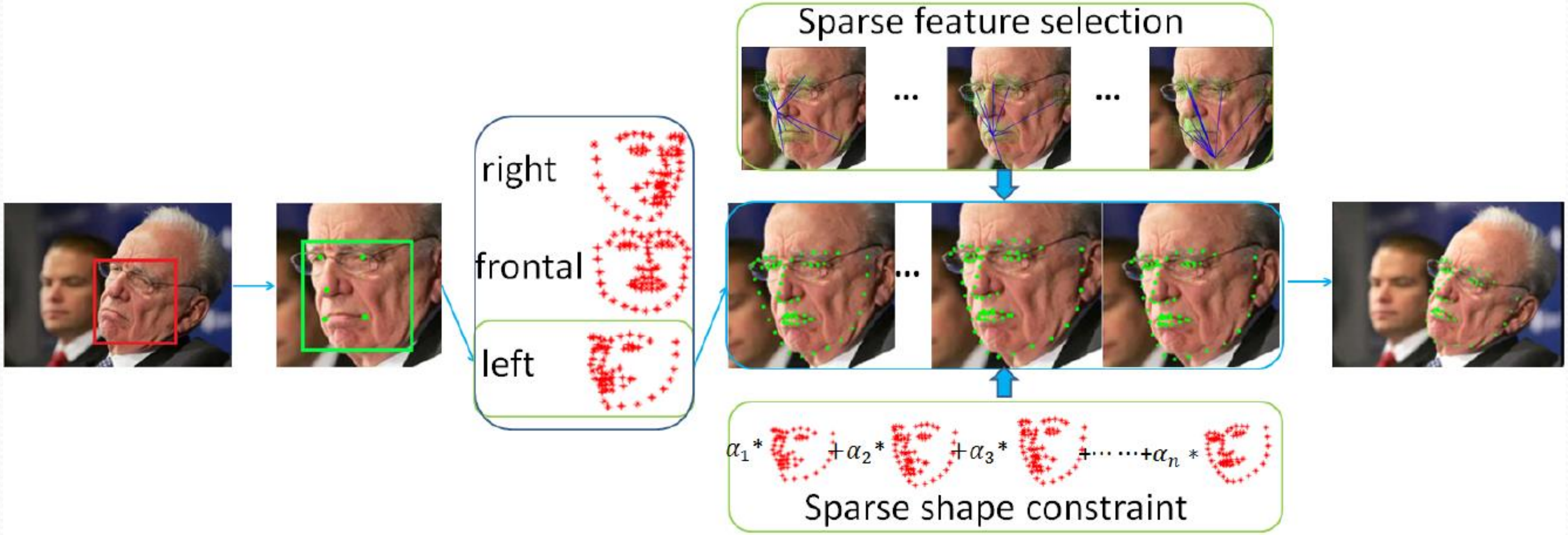
Dual sparse constrained cascade regression model (IEEE T-IP 2015)



CSR: $$\arg\min_{W_t} \sum_{i=1}^{N} \left\| \left(\boxed{X_i{}^{*}} - \boxed{X_i^{t-1}} \right) - \boxed{W_t \Phi(I_i, X_i^{t-1})} \right\|_2^2$$

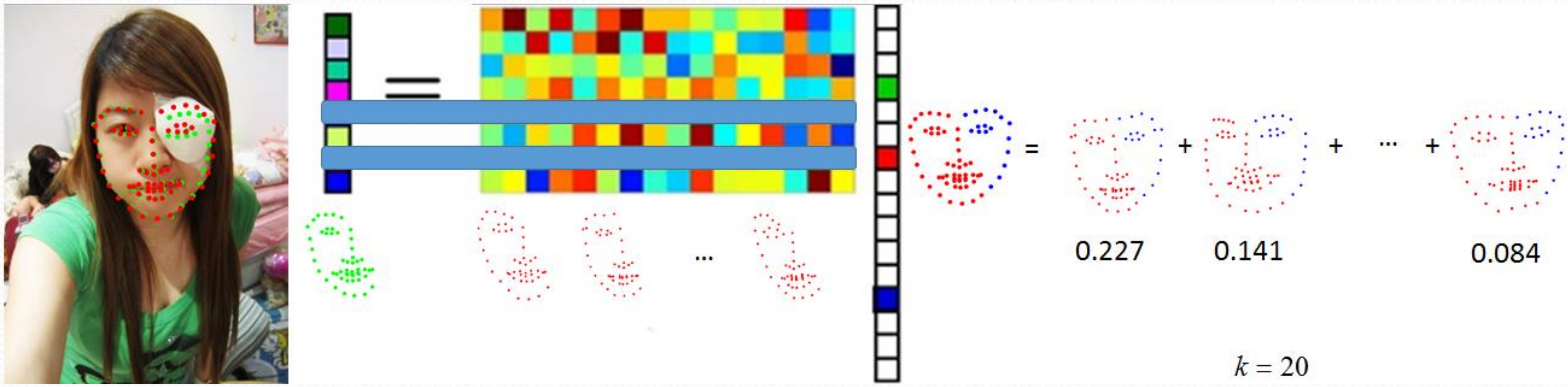
D. Piotr, W. Peter, and P. Pietro. Cascaded pose regression. *Intl. Conf. on Computer Vision and Pattern Recognition* (CVPR), 2010.

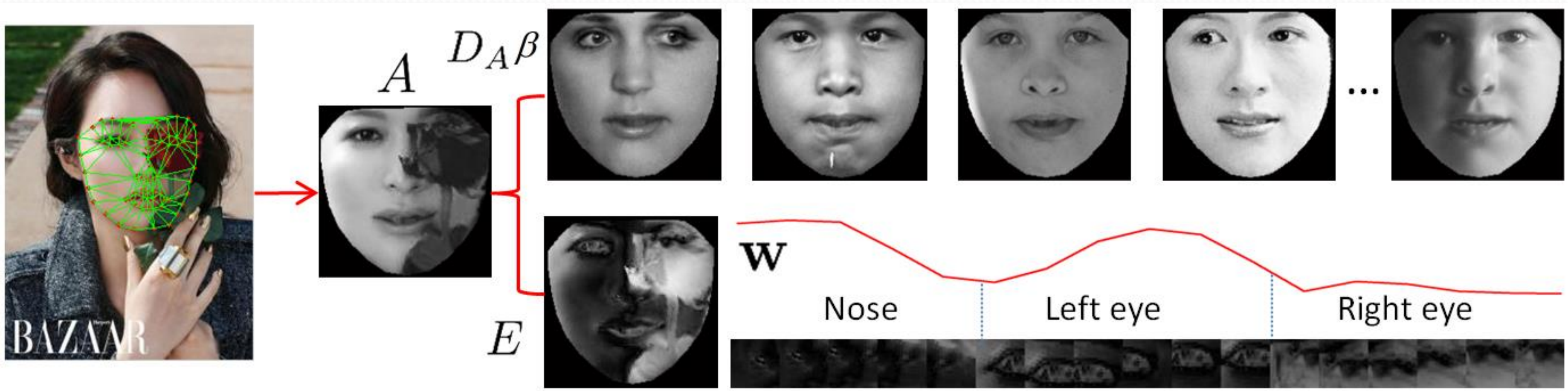
Dual sparse constrained cascade regression model (IEEE T-IP 2015)



$$\underset{\alpha,\gamma,W}{\arg\min} \left\| X^* - \Psi\left(D\alpha, \gamma\right) - W\Phi(I, \Psi\left(D\alpha, \gamma\right)) \right\|_2^2 + \lambda_1 \left\| W \right\|_1 + \lambda_2 \left\| \alpha \right\|_1$$

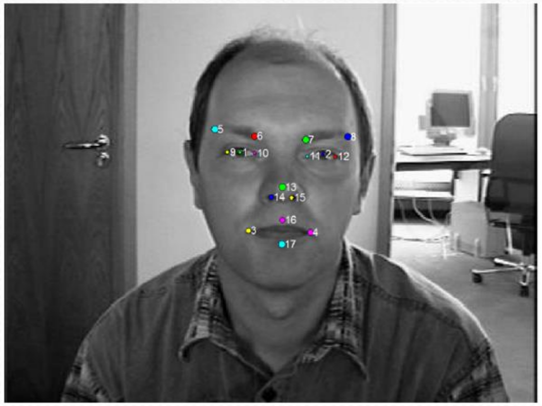
Face Alignment





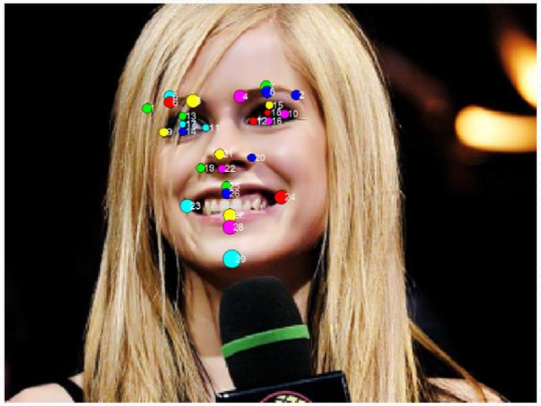
Results

LFW

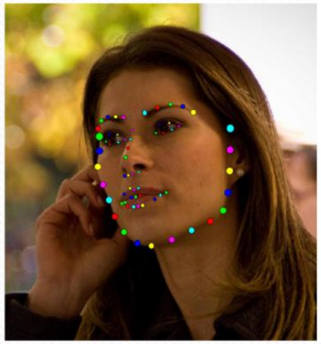


BioID

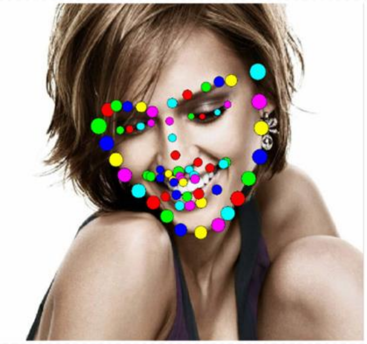
LFPW



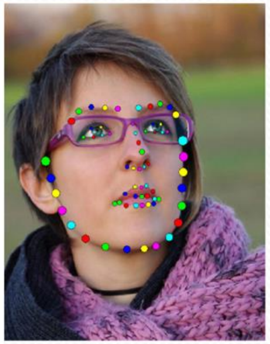
COFW



Common

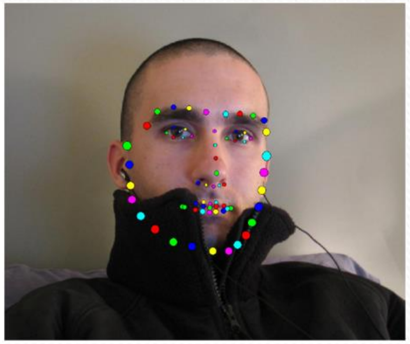


Challenge



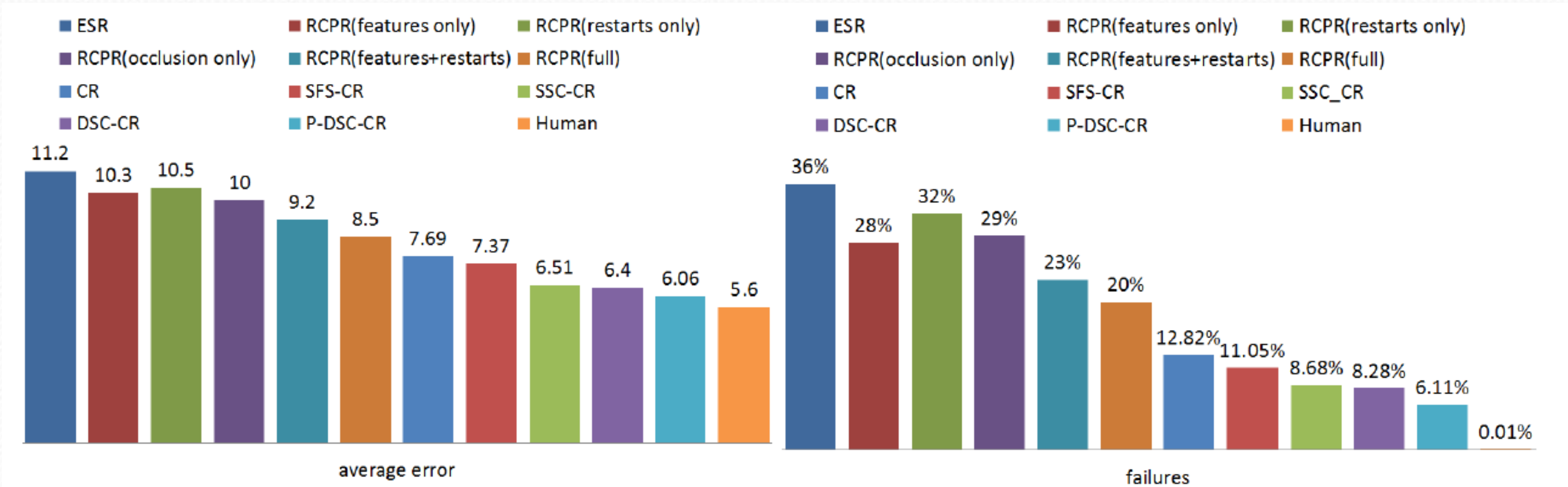
FULL

MVFW



OCFW

Results



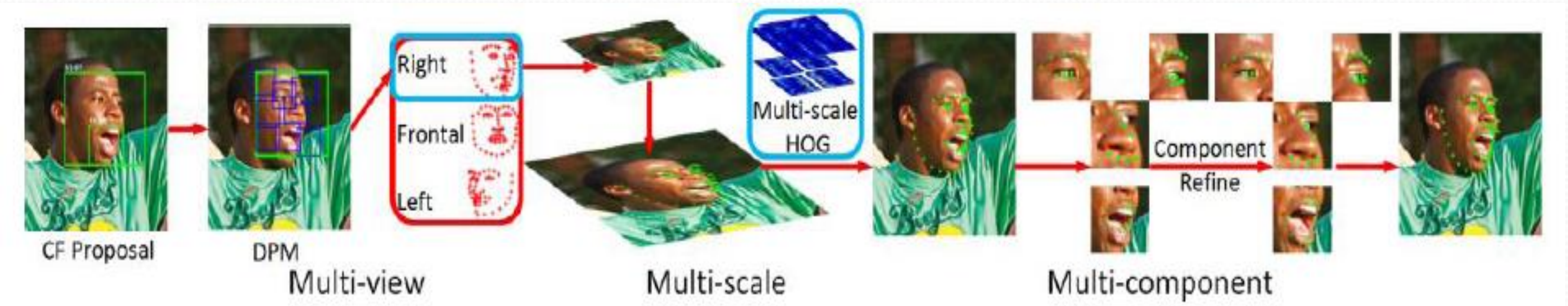
Normalized mean error and the failure rate on the COFW dataset

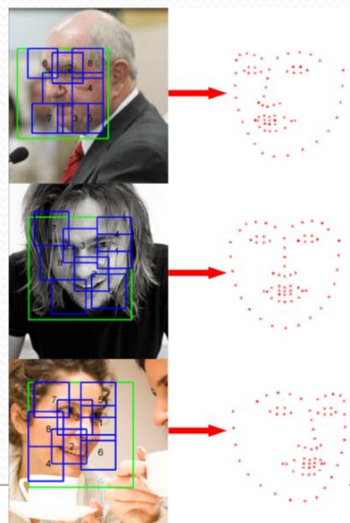
	ESR	SDM	LBF	LBF fast	TCDCN	TCDCN-Averaged	DSC-CR	P-DSC-CR
Common Subset	5.28	5.60	4.95	5.38	6.10	5.59	4.88	**3.83**
Challenging Subset	17.00	15.40	11.98	15.50	9.88	9.15	11.49	**6.93**
Fullset	7.58	7.52	6.32	7.37	6.83	6.29	6.04	**4.38**

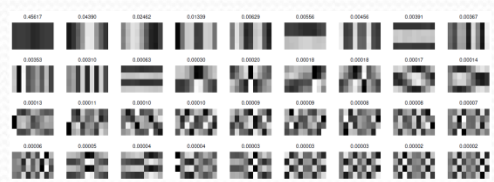
MEAN ALIGNMENT ERRORS ON THE 300-W COMMON SUBSET, CHALLENGING SUBSET AND FULLSET(*0.01)

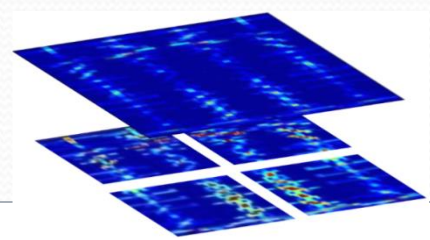
M^3 CSR model (IVC 2016)

- **Multi-view, multi-scale and multi-component**









login

intelligent behaviour understanding group

i·bug

learning - imaging - interfaces

home news projects research people publications resources vacancies courses

home » resources » 300 faces in-the-wild challenge (300-w), imavis 2014

Datasets

- Large Scale Facial Model
- 300 Videos in the Wild (300-VW) Challenge & Workshop (ICCV 2015)
- Affect "in-the-wild" Workshop
- Facial Expression Recognition and Analysis Challenge 2015
- 300 Faces In-The-Wild Challenge (300-W), IMAVIS 2014
- MAHNOB-HCI-Tagging database
- 300 Faces In-the-Wild Challenge (300-W), ICCV 2013
- MAHNOB Laughter database
- MAHNOB MHI-Mimicry database
- Facial point annotations
- MMI Facial expression database

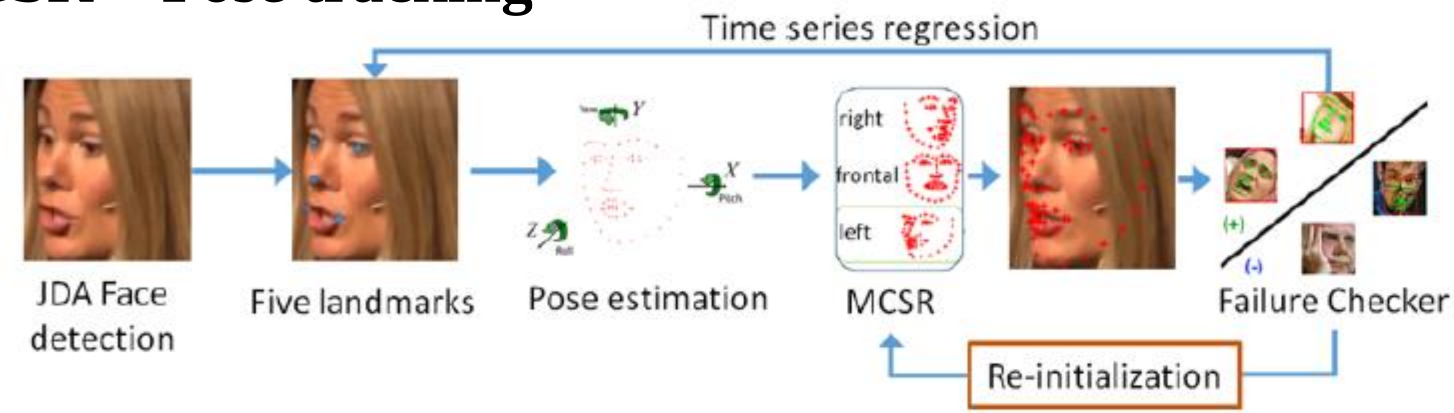
300 Faces In-The-Wild Challenge (300-W), IMAVIS 2014

Participant	*# images with detection*	*mad*		*timings (secs)*
		68 points	*51 points*	
Bongjin	584 (97.3%)	0.0271	0.0249	12.9
Y. Cheon	600 (100%)	0.1078	0.1040	0.17
J. Deng	599 (99.8%)	0.0226	0.0213	1.97
H. Fan	526 (87.7%)	0.0309	0.0294	1.29
J. Kranauskas	600 (100%)	0.0693	0.0659	2.46
S. Martin	597 (99.5%)	0.3461	0.3228	5.81
B. Martinez	600 (100%)	0.0514	0.0497	42.5
J. Shin	585 (97.5%)	0.0303	0.0287	12.6
M. Uricar	592 (98.7%)	0.0970	0.0945	3.46
F. Vojtech	591 (98.5%)	0.1047	0.0998	4.05
Oracle	–	0.0038	0.0040	–

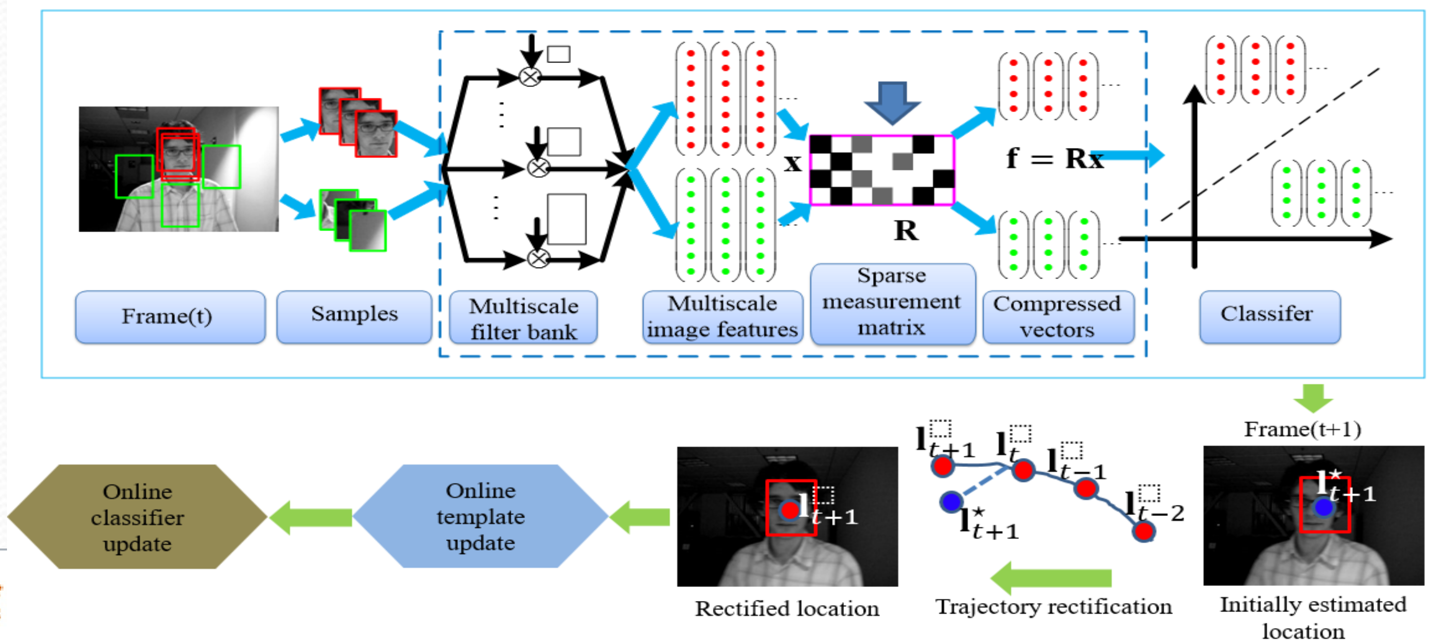
http://ibug.doc.ic.ac.uk/resources/300-W_IMAVIS/

Spatio-temporal CSR (ICCVW 2015)

- **CSR + Pose tracking**



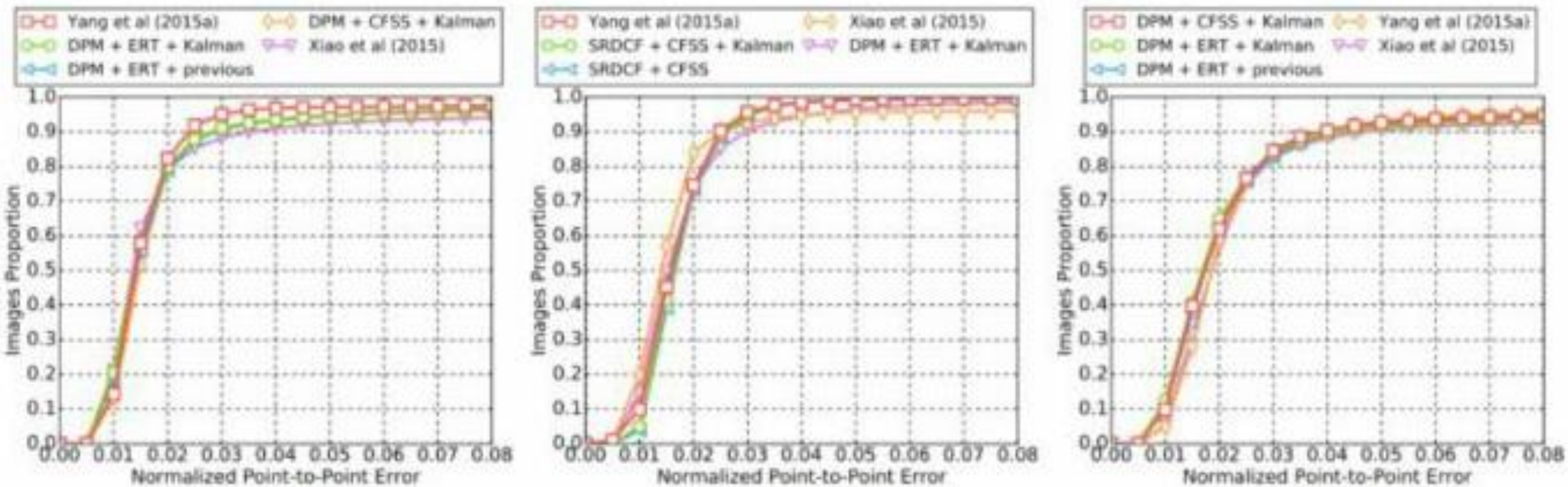
- **Adaptive compressive sensing tracker (CVIU / IEEE T-CYB 2016)**



Datasets

Large Scale Facial Model

300 Videos in the Wild (300-VW) Challenge & Workshop (ICCV 2015)

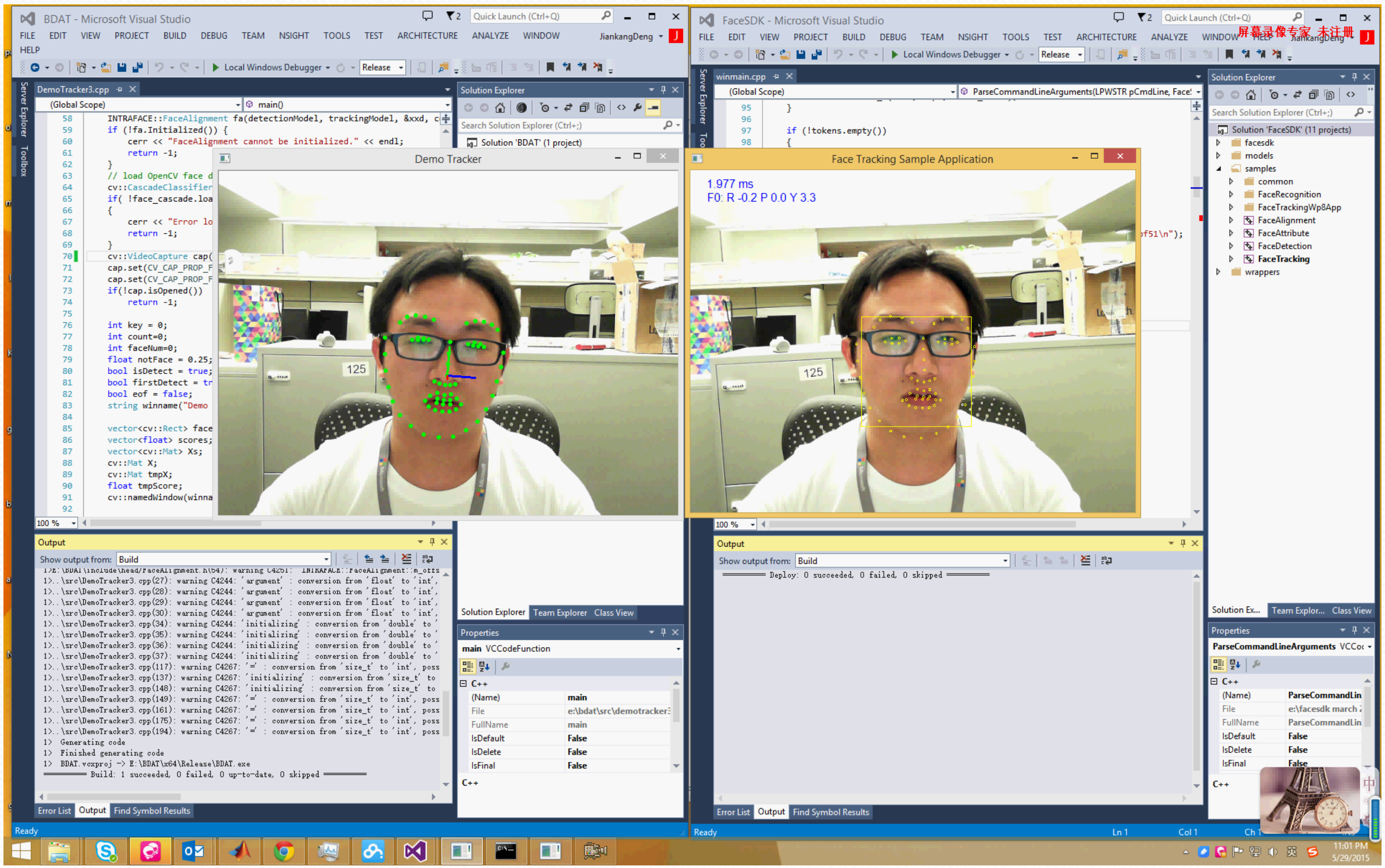


(a) Category 1 (b) Category 2 (c) Category 3

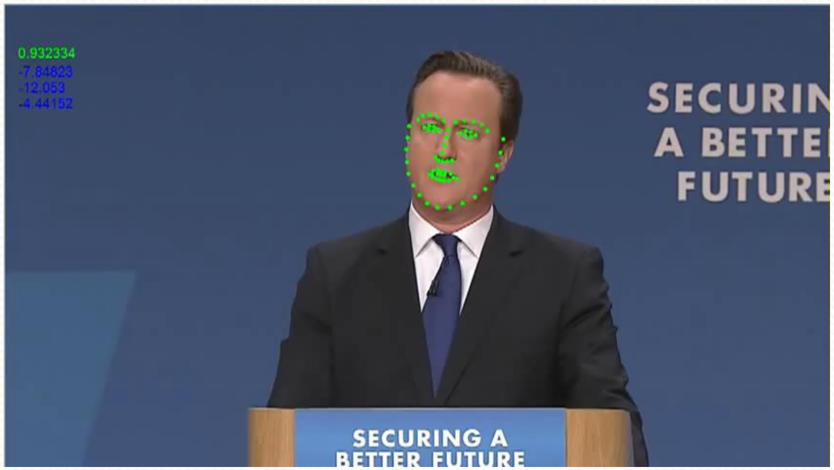
Fig. 13: Comparison between the best methods of Sections 4.3-4.7 and the participants of the 300VW challenge by Shen et al (2015). The top 5 methods are shown and are coloured red, blue, green, orange and purple, respectively. Please see Table 11 for a full summary.

curves are highlighted for each video category.

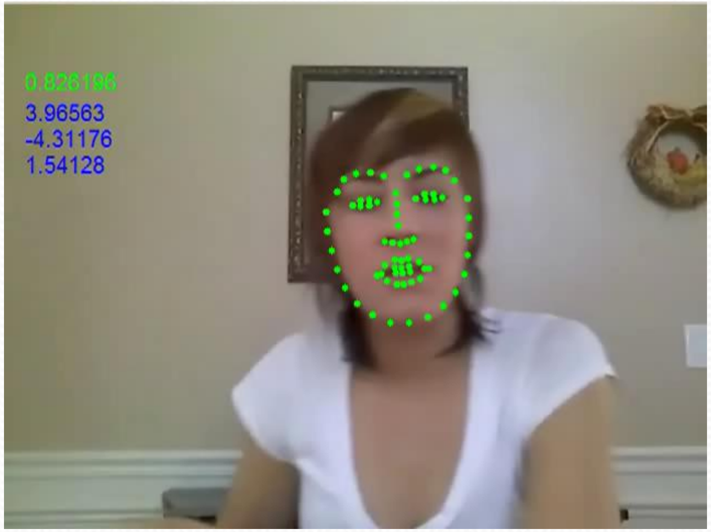
Video demo


BDAT - Microsoft Visual Studio
Demo Tracker
FaceSDK - Microsoft Visual Studio
Face Tracking Sample Application
1.977 ms
F0: R -0.2 P 0.0 Y 3.3

Video demo


0.932334
-7.84823
-12.053
-4.44152
SECURIN
A BETTE
FUTURE
SECURING A
BETTER FUTURE


RDC
FILMU
dr Magdalena OGÓREK
opozycja tezy


0.826196
3.96563
-4.31176
1.54128


0.905905
9.22949
-10.3213
-5.00254
-...sì.
-...in condizioni terribili.

Live demo

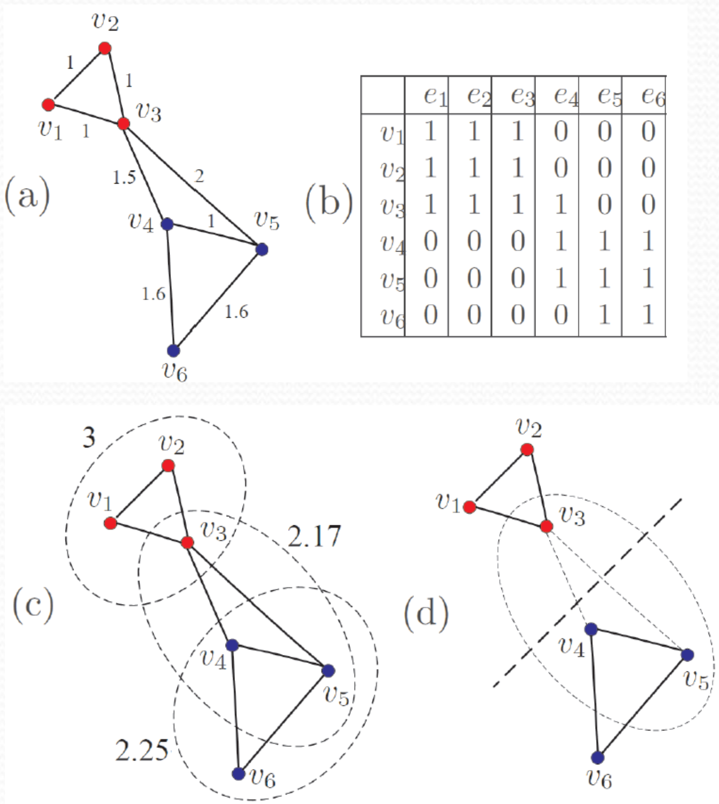

HUAWEI

Why is hypergraph?

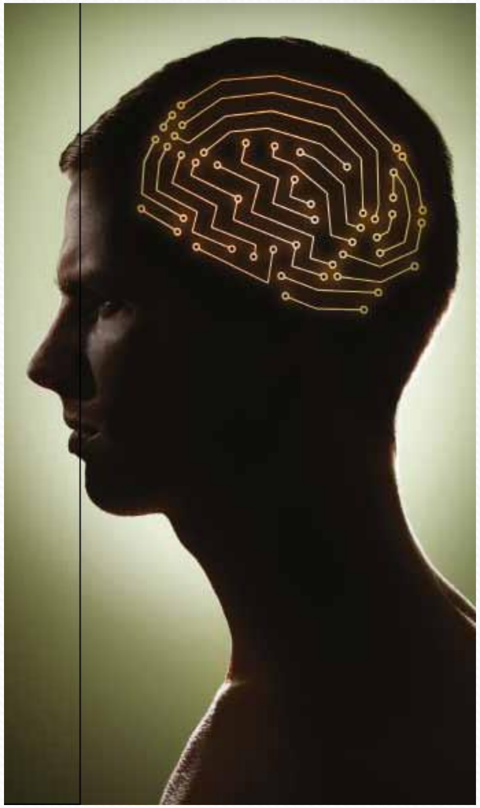
Six images fromCaltech-101. The first three images are from the 'ferry' class; the last three images are from the 'joshua tree' class.

How to build the complicated relationship of multiple features?

Why is hypergraph?

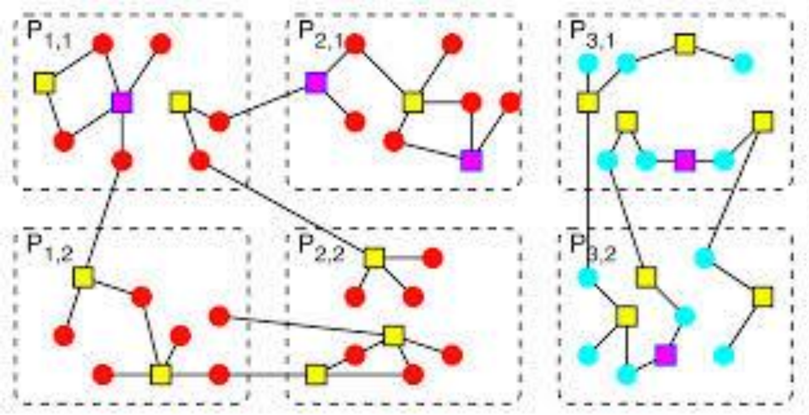


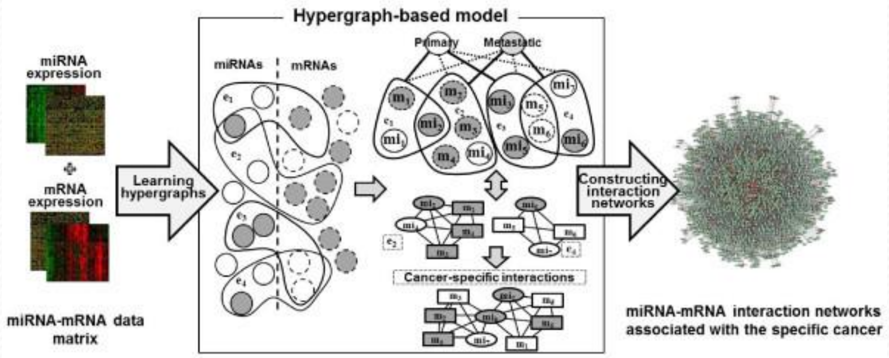
	e_1	e_2	e_3	e_4	e_5	e_6
v_1	1	1	1	0	0	0
v_2	1	1	1	0	0	0
v_3	1	1	1	1	0	0
v_4	0	0	0	1	1	1
v_5	0	0	0	1	1	1
v_6	0	0	0	0	1	1



- It is not complete to represent the relations among vertices only by pairwise simple graphs.
- It may be helpful to take account of the relationship not only between two vertices, but also among three or more vertices containing local grouping information.

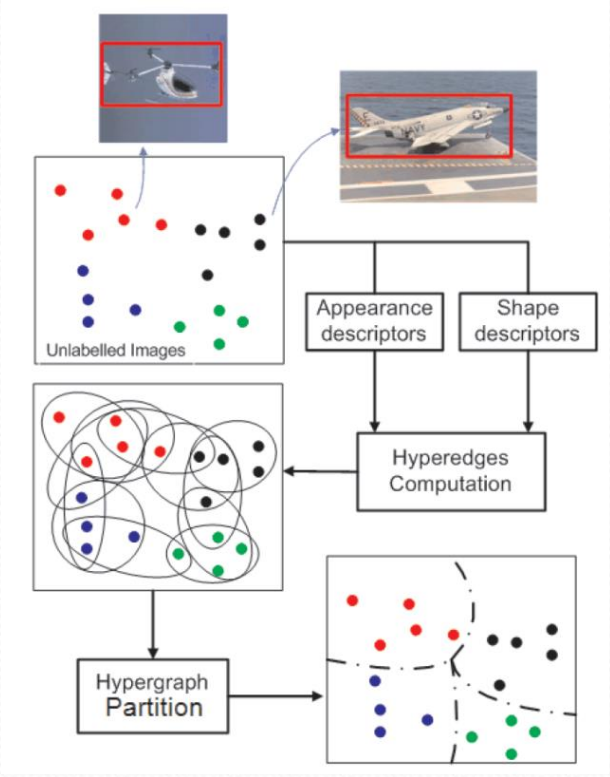
Why is Hypergraph?



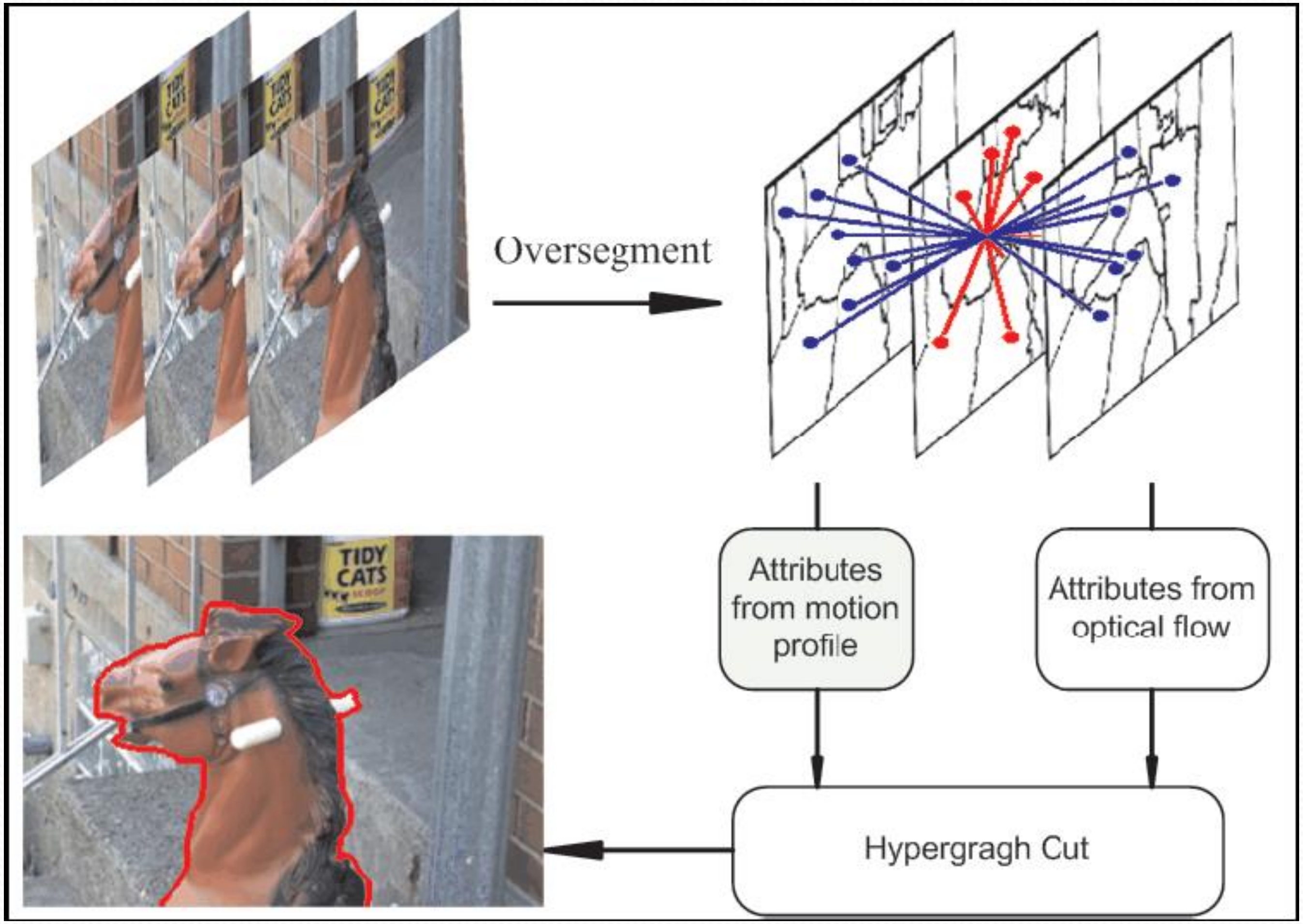


Hypergraph-based feature representation

- Unsupervised hypergraph learning
  - Video objects clustering (CVPR 2009)
  - Image categorization (TPAMI 2011)
- Semi-supervised hypergraph learning
  - Content-based image retrieval (CVPR 2010, PR 2011)
- Sparse hypergraph learning
  - Elastic hypergraph (TIP 2016)
  - Application in hyperspectral image classification (TGRS submitted)



Video Object Segmentation (ICCV 2009)



Results-Squirrel

Ground Truth

Simple Graph + Optical Flow

Simple Graph + Motion Profile

Simple Graph + Both Motion Cues

Hypergraph Cut

Results-Walking with Rotation

Ground Truth

Simple Graph + Optical Flow

Simple Graph + Motion Profile

Simple Graph + Both Motion Cues

Hypergraph Cut

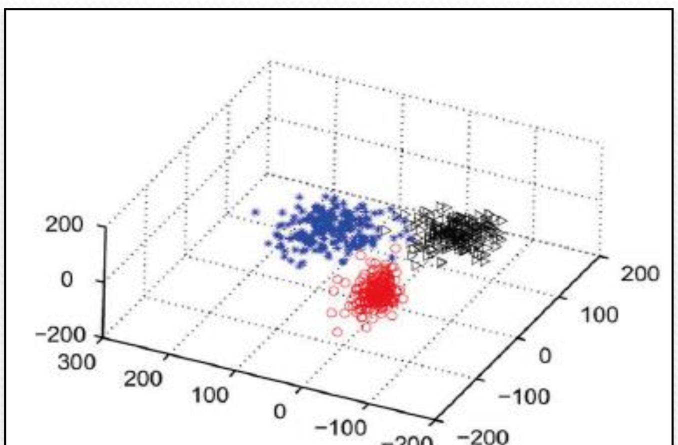
Videos

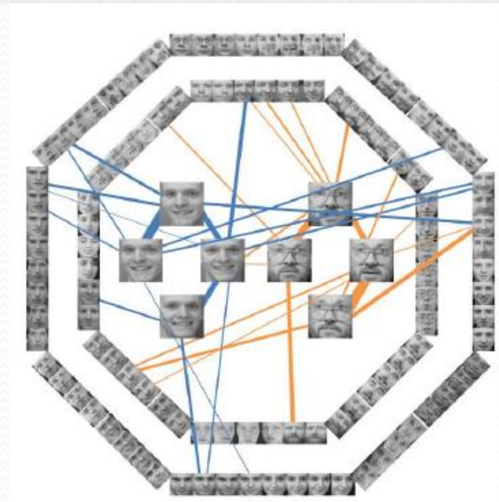
Elastic Net Hypergraph Learning (IEEE T-IP 2016)

Robust Elastic Net Representation

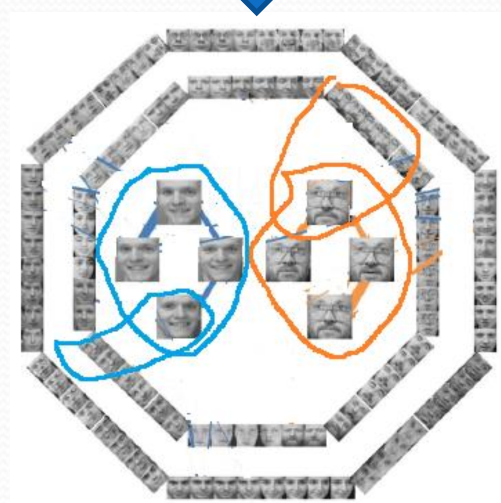
$$\min_{Z,S} \|Z\|_1 + \lambda\|Z\|_F^2 + \gamma\|S\|_{2,1}$$
$$s.t. X = XZ + S, diag(Z) = 0$$

Hypergraph Learning





KNN-Graph



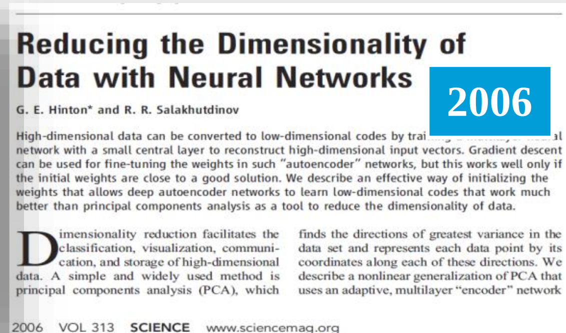
Elastic Net Hypergraph

Elastic Net Hypergraph Learning (IEEE T-IP 2016)

Dataset	G-graph	LE-graph	l_1-graph	KNN-HG	SCHG	l_1-Hypergraph	ENHG
Extended Yale B (10%)	66.49	70.79	53.87	71.80	77.68	82.15	**88.59**
Extended Yale B (20%)	65.34	69.97	54.46	75.54	81.80	83.48	**90.87**
Extended Yale B (30%)	33.72	71.85	53.90	77.67	82.84	85.36	**93.94**
Extended Yale B (40%)	66.28	71.34	56.61	80.59	83.55	86.90	**94.34**
Extended Yale B (50%)	66.90	71.60	57.75	80.80	84.48	87.08	**94.97**
Extended Yale B (60%)	67.52	71.48	58.48	81.79	89.46	90.42	**95.28**
PIE (10%)	65.72	67.75	78.29	68.74	79.35	80.24	**88.31**
PIE (20%)	66.94	69.58	82.82	70.18	84.74	84.55	**94.94**
PIE (30%)	69.89	73.48	87.94	74.39	88.78	89.29	**96.55**
PIE (40%)	71.54	76.38	90.99	76.14	90.33	91.75	**97.33**
PIE (50%)	73.04	78.35	93.39	78.76	92.66	93.71	**97.53**
PIE (60%)	74.91	80.44	95.00	79.95	94.12	94.87	**98.43**
USPS (10%)	96.87	96.79	88.33	96.51	97.08	97.20	**97.36**
USPS (20%)	97.78	97.90	91.11	98.17	98.12	98.29	**98.39**
USPS (30%)	98.45	98.47	93.08	98.78	98.87	98.85	**98.91**
USPS (40%)	98.80	98.82	95.96	99.08	99.08	**99.10**	99.08
USPS (50%)	99.18	99.14	97.31	99.39	**99.41**	99.39	99.40
USPS (60%)	99.35	99.28	98.86	99.51	99.50	99.52	**99.53**

IMAGENET

Deep Learning



Reducing the Dimensionality of Data with Neural Networks

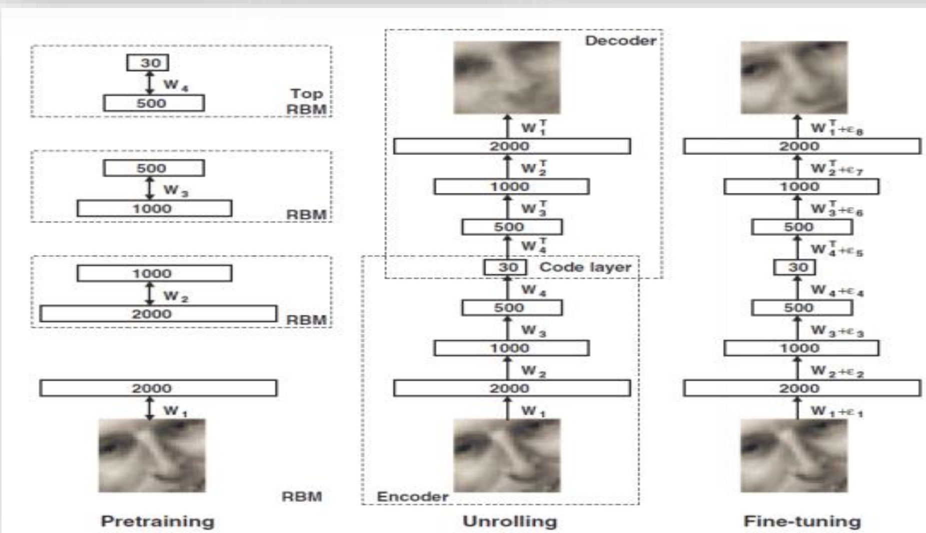
G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by trai... 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.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

2006 VOL 313 SCIENCE www.sciencemag.org

2006



Breakthrough

2011年 Speech recogniton

Microsoft

task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

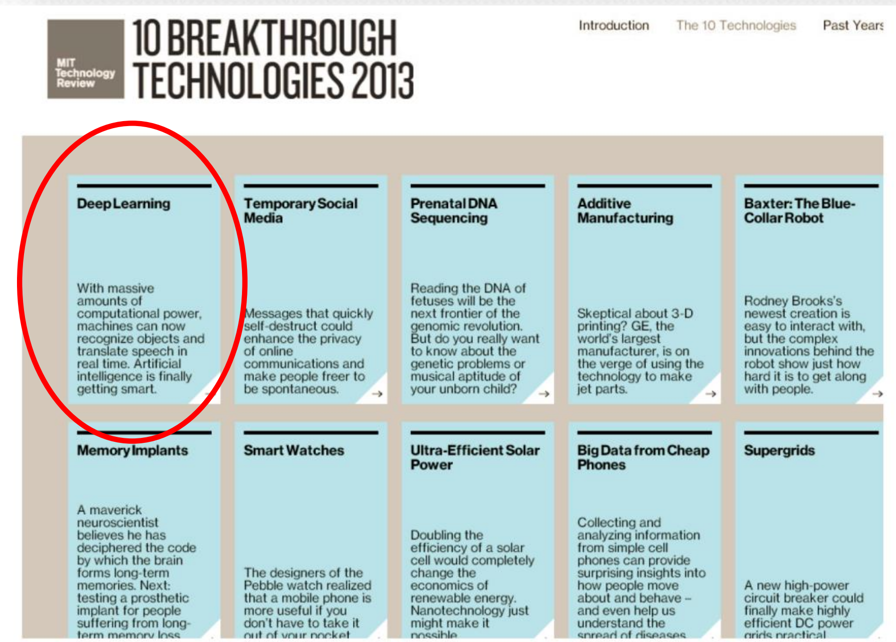
2012年 Image classification

IMAGENET

Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

IMAGENET

No. 1 in 10 breakthrough tech 2013 selected by MIT tech review



MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Introduction The 10 Technologies Past Years

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child?

Additive Manufacturing

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts.

Baxter: The Blue-Collar Robot

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people.

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss.

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket.

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible.

Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave – and even help us understand the spread of diseases.

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical.

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

2015年5月Nature杂志以综述的形式对深度学习进行了总结和评价，指出深度学习最大的优点是能自动学习和抽象数据特征

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer typically detects motifs by spotting particular arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects

Supervised learning

The most common form of machine learning, deep or not, is supervised learning. Imagine that we want to build a system that can classify images as containing, say, a house, a car, a person or a pet. We first collect a large data set of images of houses, cars, people and pets, each labelled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training. We compute an objective function that measures the error (or distance) between the output scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters, often called weights, are real numbers that can be seen as 'knobs' that define the input–output function of the machine. In a typical deep-learning system, there may be hundreds of millions of these adjustable weights, and hundreds of millions of labelled examples with which to train the machine.

IMAGENET Large Scale Visual Recognition Challenge 2015 (ILSVRC2015)

News H... ...rces Registration FAQ Citation Contact

News

- F...
- J...
- J...
- D...
- D...
- N...
- N...
- O... ...announced.
- O...
- August 15, 2015: Regis...
- August 13, 2015: Comp...
- June 12, 2015: Tentativ...
- June 2, 2015: Additiona...
- May 19, 2015: Announc...
- December 17, 2015: St...
- September 19, 2014: Tr... ...NC Chapel Hill.

History

2014, 2013, 2012, 2011, 2010

Task 1b: Object detection with additional training data

Ordered by number of categories won

Team name	Entry description	Description of outside data used	**Number of object categories won**	mean AP
Amax	remove threshold compared to entry1	pre-trained model from classification task;add training examples for class number <1000	165	0.57848
CUImage	Combined models with region proposals of cascaded RPN, edgebox and selective search	3000-class classification images from ImageNet are used to pre-train CNN	30	0.522833
MIL-UT	ensemble of 4 r...			
Amax	Cascade region...			
MIL-UT	ensemble of 4 r... learned separat...			

Task 3b: Object detection from video with additional training data

Ordered by number of categories won

Team name	Entry description	Description of outside data used	**Number of object categories won**	mean AP
Amax	only half of the videos are tracked due to deadline limits, others are only detected by Faster RCNN (VGG16) without tempor smooth.	---	18	0.730746
CUVideo	Outside training data (ImageNet 3000-class data) to pre-train the detection model, mAP 77.0 on validation data	ImageNet 3000-class data to pre-train the model	11	0.696607
Trimps-Soushen	Models combine with m... constraint			
Trimps-Soushen	Combine several mode...			
BAD	VID2015_trace_merge...			
BAD	combined_VIDtrainval_...			
BAD	VID2015_merge_test_t...			
BAD	combined_test_DET_th...			
BAD	VID2015_VID_test_thre...			

Task 2b: Classification+localization with additional training data

Ordered by classification error

Team name	Entry description	Description of outside data used	**Classification error**	Localization error
Amax	Validate the classification model we used in DET entry1	share proposal procedure with DET for convinence	0.04354	0.14574
Trimps-Soushen	extra annotations collected by ourselves	extra annotations collected by ourselves	0.04581	0.122285
CUImage	Average multiple models. Validation accuracy is 79.78%.	3000-class classification images from ImageNet are used to pre-train CNN	0.05858	0.198272

IMAGENET

IMAGENET Large Scale Visual Recognition Challenge 2016 (ILSVRC2016)

News History

News

- Septemb
- Septemb
- Septemb
- Septemb
- May 31, :
- May 26, 2016:
- May 3, 2016: S

er 18, 2016 5pm PST.

History

2015, 2014, 2013, 201

Tentative Time

- May 31, 2016:
- September 9, 2016, !

Object detection from video (VID)[top]

Task 3a: Object detection from video with provided training data

Ordered by number of categories won

Team name	Entry description	**Number of object categories won**	mean AP
NUIST	cascaded region regression + tracking	10	0.808292
NUIST	cascaded region regression + tracking	10	0.803154
CUVi	4-model ensemble with Multi-Context Suppression and Motion-Guided		

Task 3b: Object detection from video with additional training data

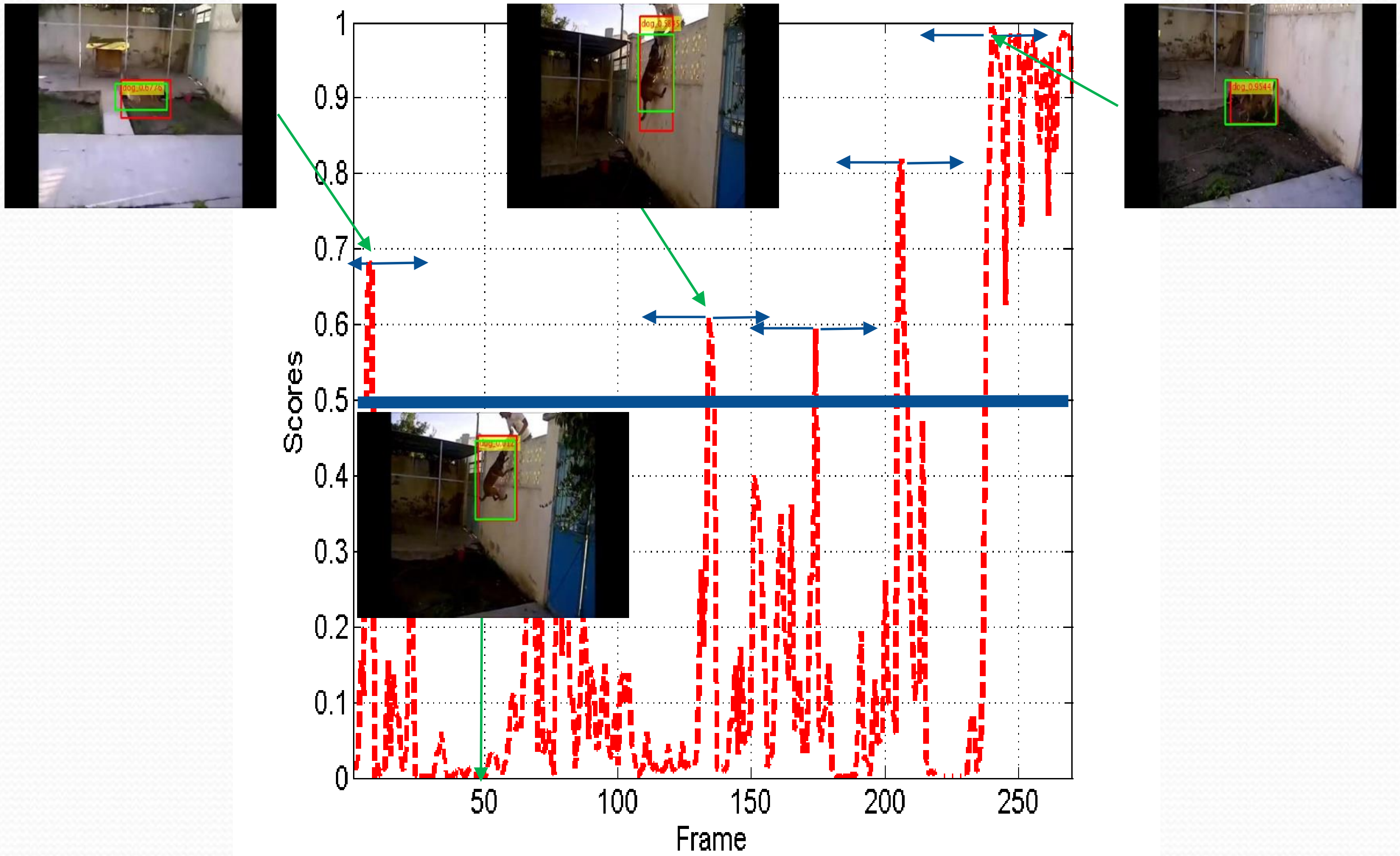
Ordered by number of categories won

Team name	Entry description	Description of outside data used	**Number of object categories won**	mean AP
NUIST	cascaded region regression + tracking	proposal network is finetuned from COCO	17	0.79593
NUIST	cascaded region regression + tracking	proposal network is finetuned from COCO	5	0.781144
Trimps-Sous		Extra data from ImageNet dataset(out of the		

Task 3d: Object detection/tracking from video with additional training data

Team name	Entry description	Description of outside data used	**mean AP**
NUIST	cascaded region regression + tracking	proposal network is finetuned from COCO	0.583898
ITLab-	An ensemble for detection,	pre-trained model from COCO detection, extra data collected	0.490863

Object detection from Video

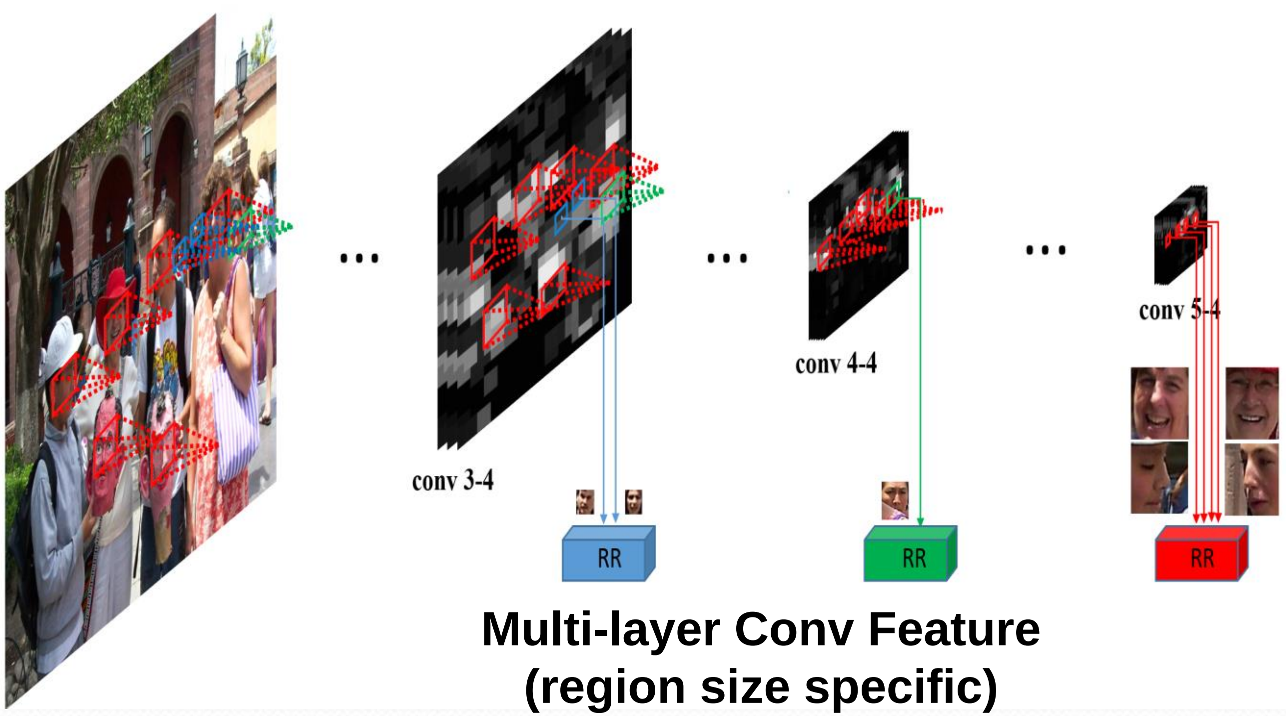


Object detection on each frame

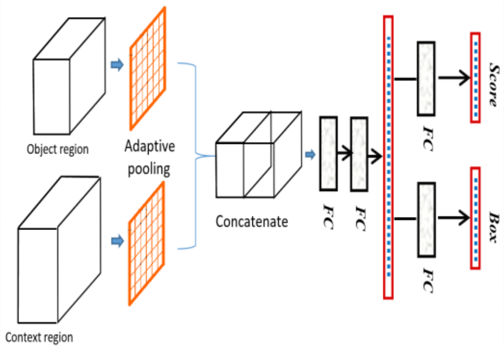
Tracking from the high score frame (temporal smooth)

Class-wise box regression and NMS on each frame

Cascade Region Regression



Multi-layer Conv Feature (region size specific)

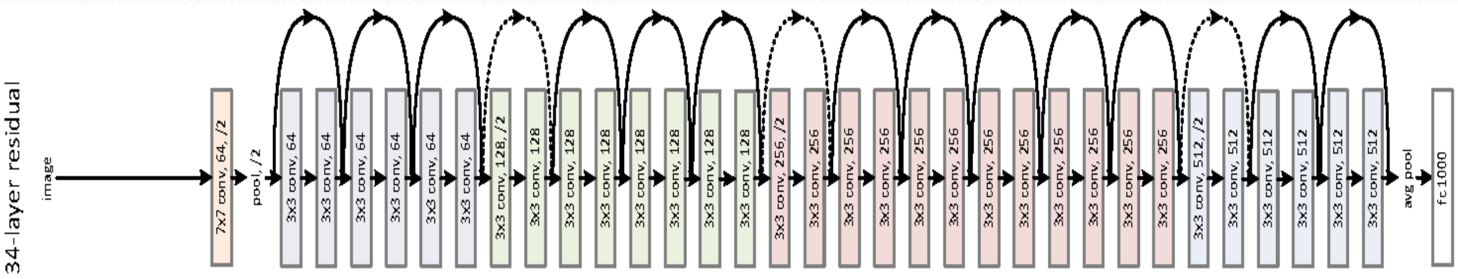


Multi-scale Conv Feature (object + around context)

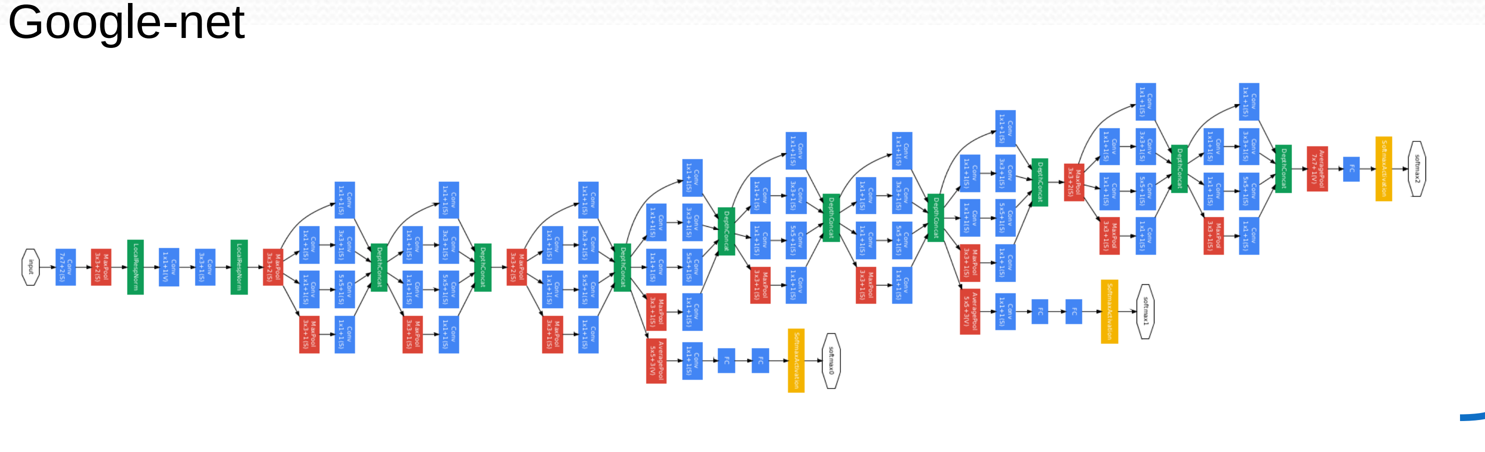
Cascade region regression根据region的大小选择不同层的region regressor对bounding box进行调整，较大的region使用后面的feature map，较小的feature map使用前面的feature map。

Model Ensemble

Res-net



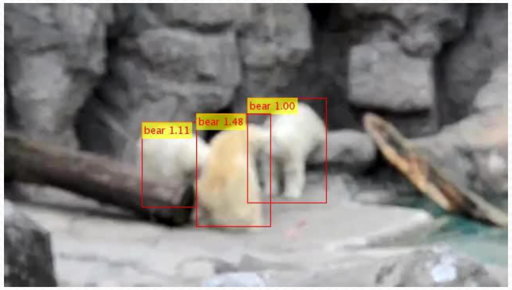
Google-net



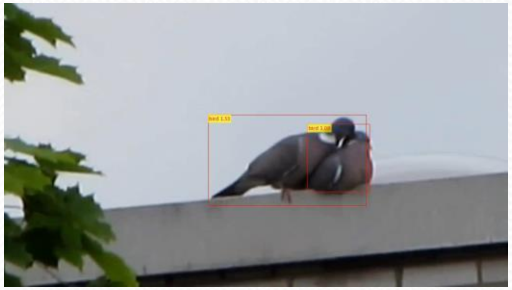
Model ensemble is always effective.

Demo Video


squirrel 1.83
Red Squirrel

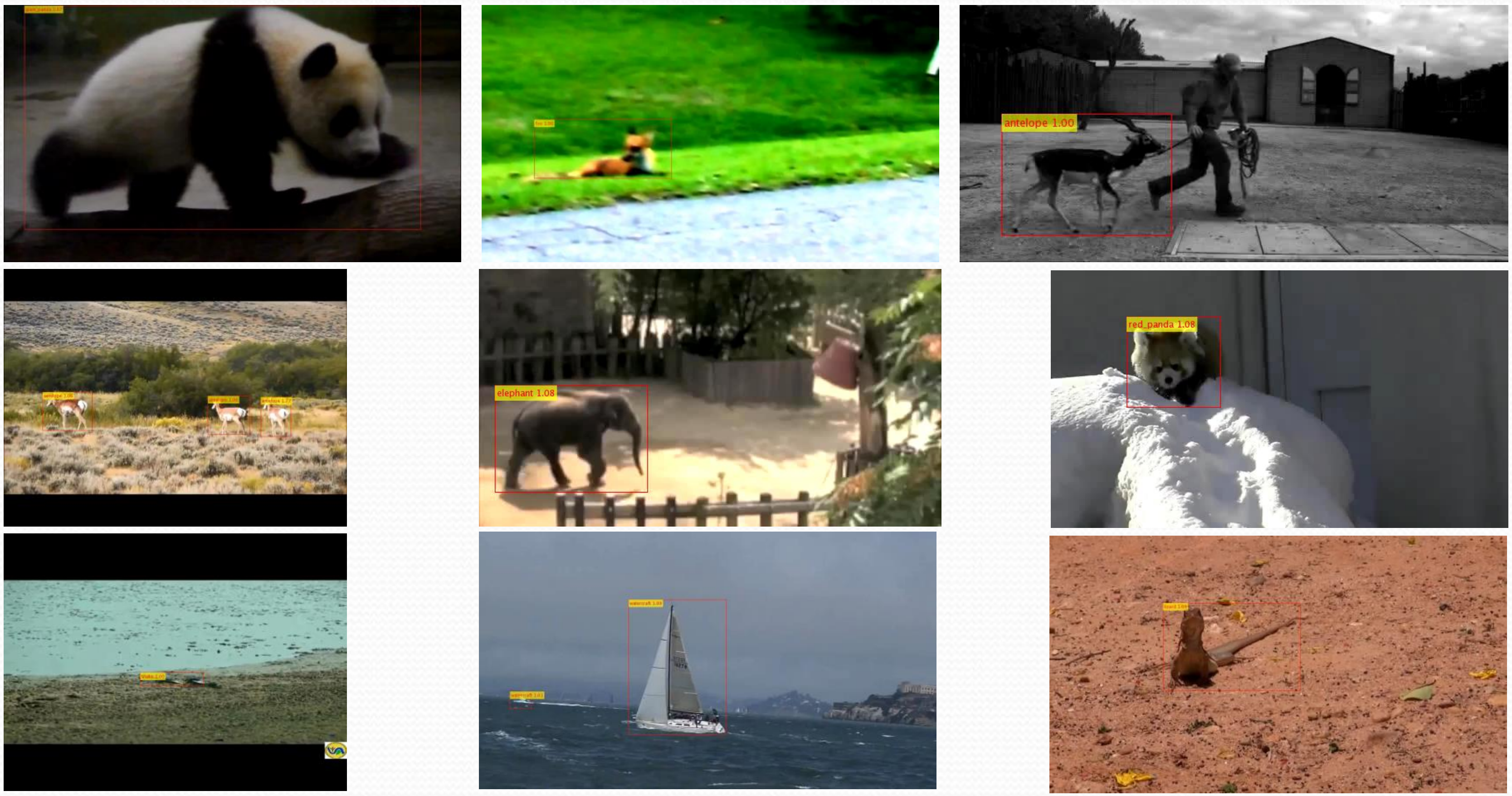

bear 1.11
bear 1.48
bear 1.00

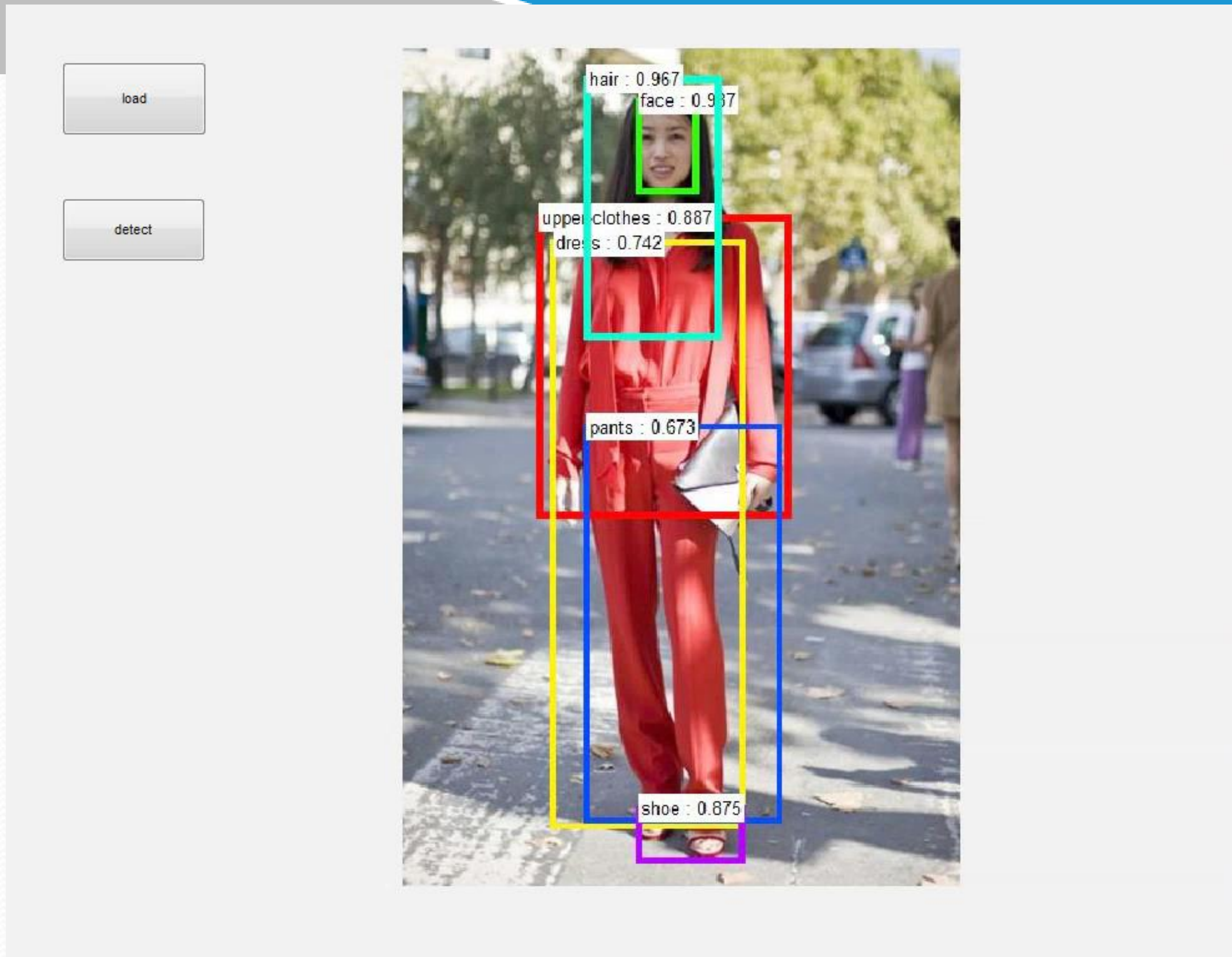

elephant 1.23
elephant 1.14
elephant 1.14

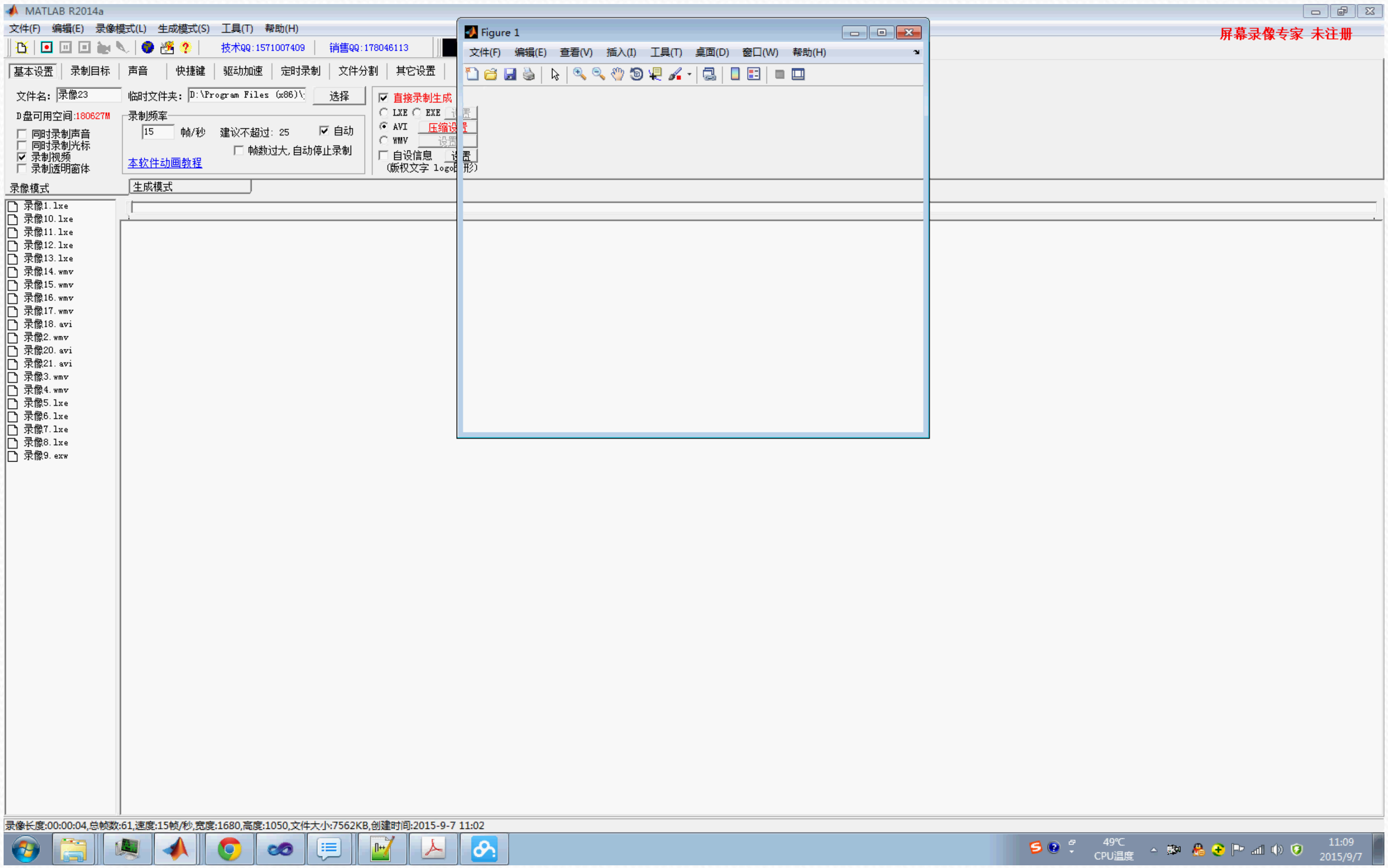



domestic_cat 1.17

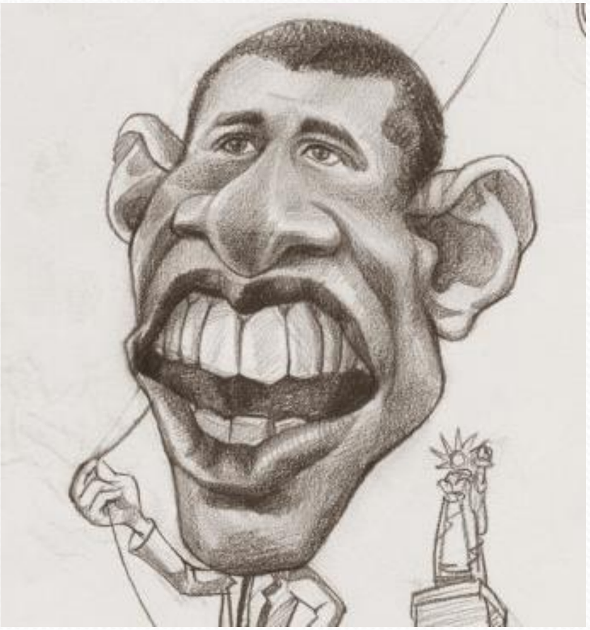
Demo Video

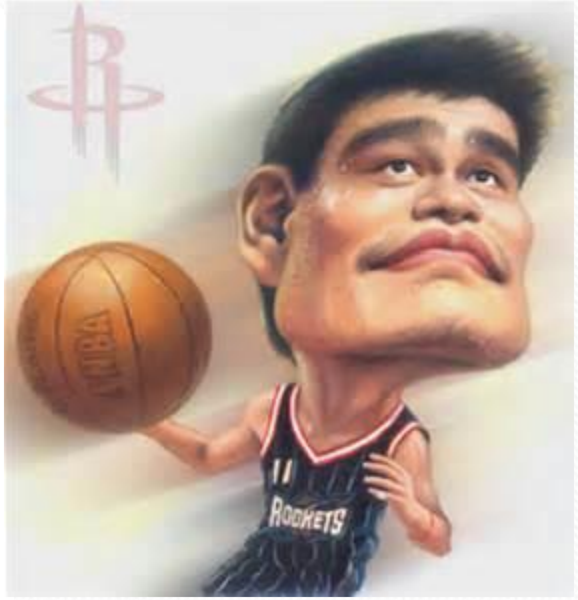



load
detect
hair : 0.967
face
upper clothes : 0.887
dress : 0.742
pants : 0.673
shoe : 0.875


MATLAB R2014a
屏幕录像专家 未注册
Figure 1
录像长度:00:00:04,总帧数:61,速度:15帧/秒,宽度:1680,高度:1050,文件大小:7562KB,创建时间:2015-9-7 11:02

Does Cartoonist use deep features?





Thank You

Qingshan Liu

Email: qsliu@nuist.edu.cn
Cell: 13585199482