中国计算机学会计算机视觉专委会走进高校系列报告会

中国石油大学（华东）·青岛（第25期）

面向大规模场景与行为分类的深度学习技术

Deep learning methods for large scale scene and action recognition

乔宇

中国科学院深圳先进技术研究院

2016-Dec-16

Outline

- Overview
- Deep learning for large scale scene classification
  - Knowledge Guided Disambiguation for Scene Recognition with Multi-Resolution CNNs
- Deep learning approaches for action recognition
  - Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors (CVPR 2015)
  - Actioness estimation (CVPR 2016)
  - Real-time Action Recognition with Enhanced Motion Vector CNNs (CVPR 2016)
  - Deep Segmental Network (NO 1 in ActivityNet 2016)
- Conclusions

深度学习方法快速发展

深度神经网络已经在语音、视觉、自然语言处理等领域取得了很大成功，在学术界和工业界都引起了极大关注。

深度学习理论突破

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 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.

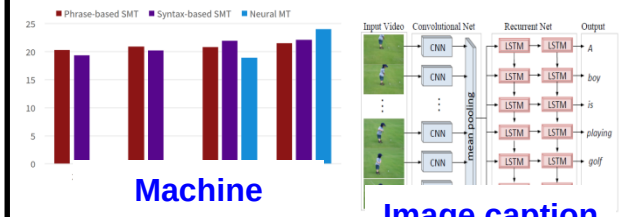
Dimensionality reduction facilitates the classification, visualization, communi- finds the directions of greatest variance in the data set and represents each data point by its

深度置信网络

ImageNet竞赛: 74% vs. 85%

1000类，1百万数据

机器翻译和文本生成

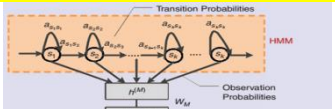


Machine Translation

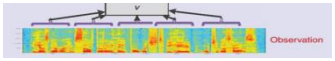
Image caption

2006 2011 2012 2013 2015 2016

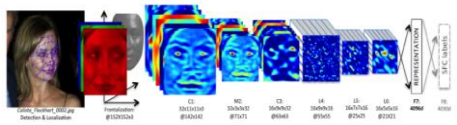
大规模语音识别



Switchboard：错误率降低8.9%



人脸识别



在LFW上识别率99%,超过人类

围棋

Alphago 4:1 李世乭

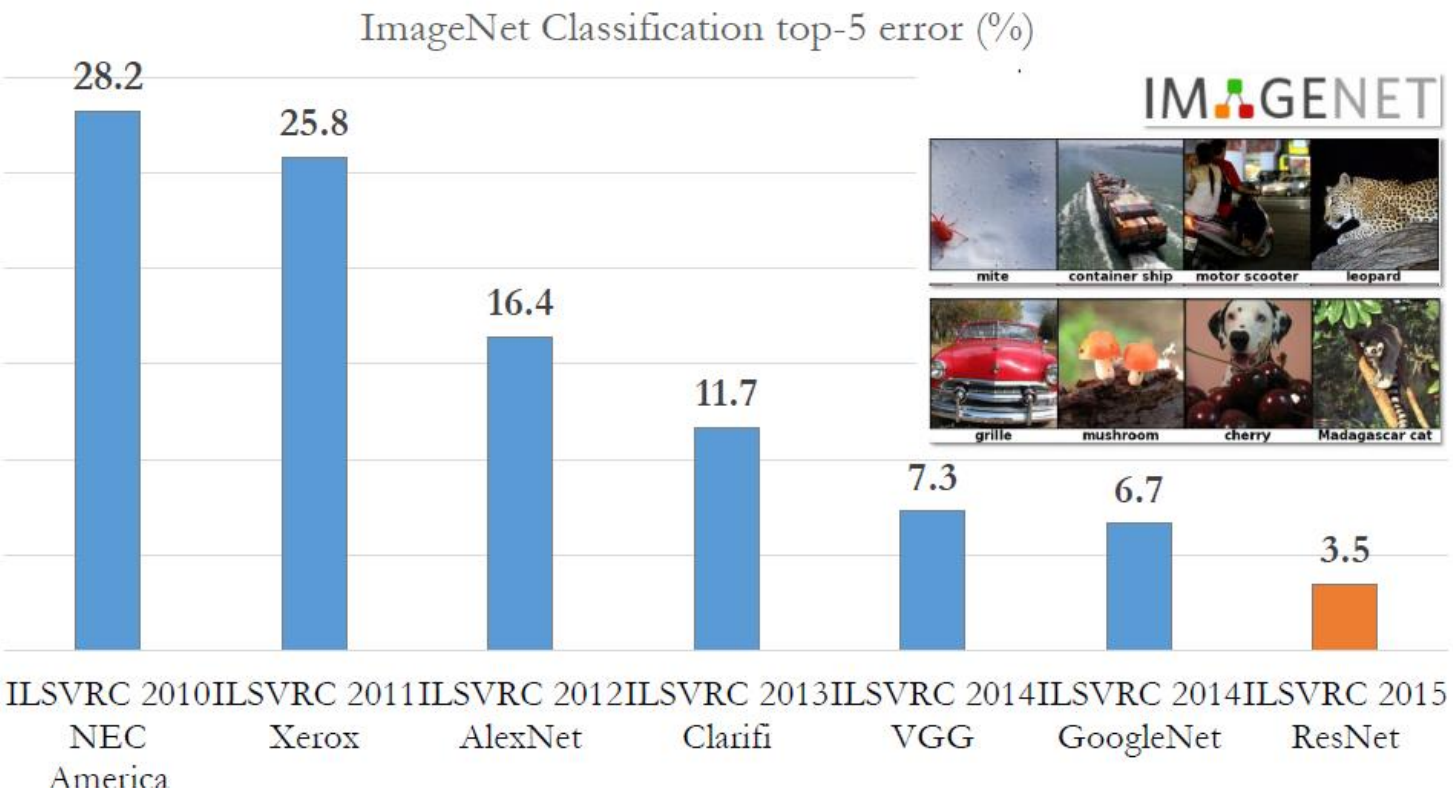
ImageNet 数据库

ImageNet 包含100万图像，1000个类别。

Large Scale Visual Recognition Challenge from 2010

http://www.image-net.org/

ImageNet 1000类图像分类



ImageNet竞赛结果

深度学习方法
非深度学习方法

Top 5 error rates in testing data

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

MS 3.57%, Google 3.46% in 2015。人的误差是5.1%

ImageNet 2012 第一名AlexNet

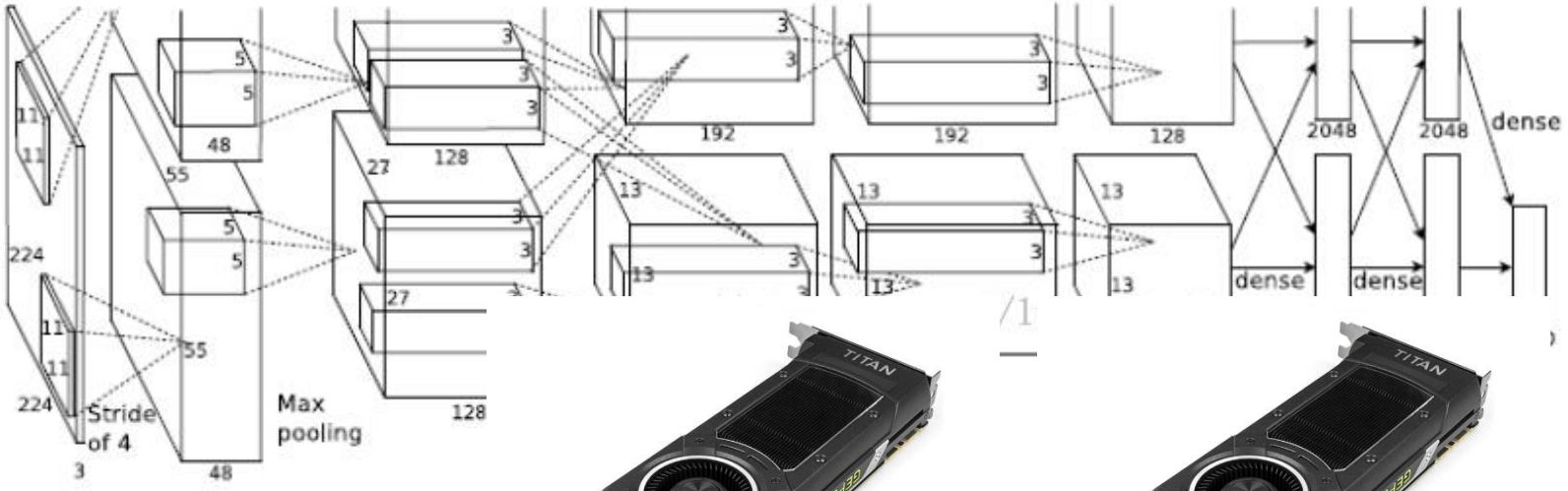
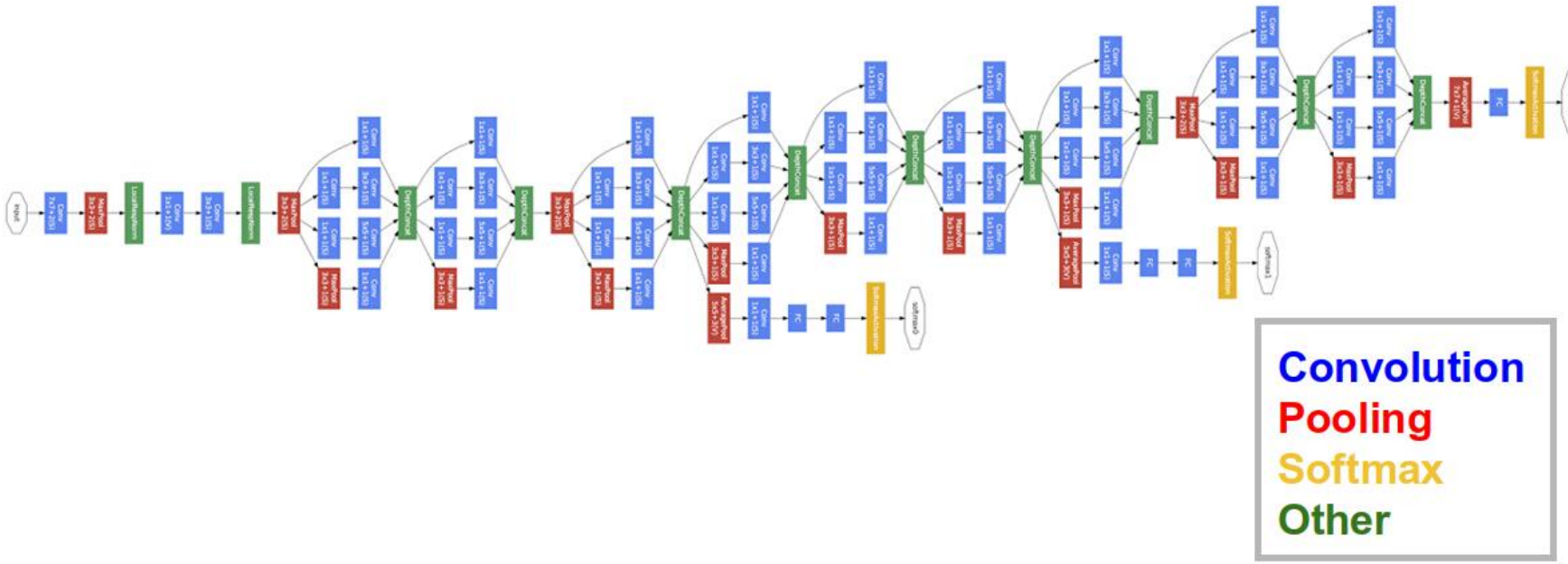


Figure 2: An illustration of the ar ow s
between the two GPUs. One GPU igu s
at the bottom. The GPUs commun or d
the number of neurons in the netwo
4096–4096–1000.

From: (2012_NIPS) Imagenet classification with deep convolutional neural networks.

AlexNet7层，2400万节点，1亿4千万参数和150亿联结。ImageNet top5错误率 15.3%

ImageNet 2014 第一名GoogLeNet

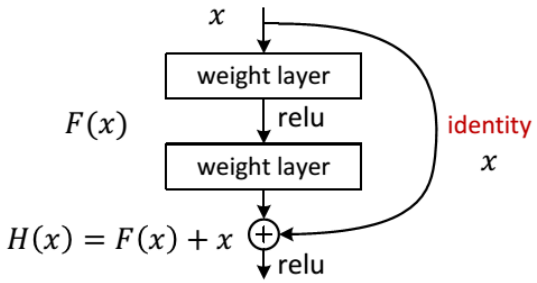


GoogLeNet 22层, ImageNet top5错误率6.6%。

《Going Deeper with Convolutions》, Arixv 2014

ImageNet 2015 第一名ResNet

- Residual net



MSRA team

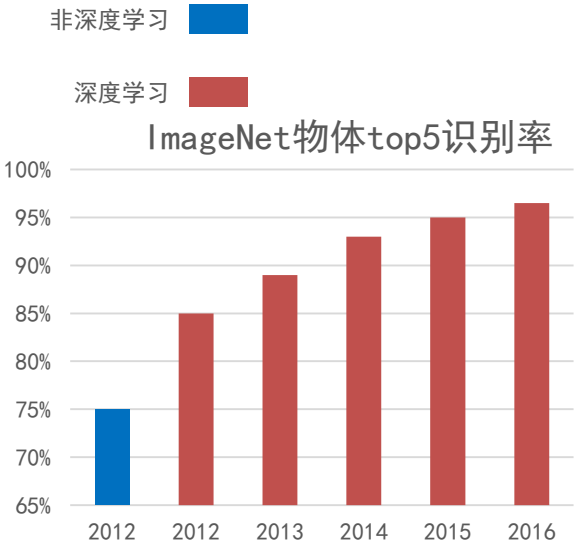
ResNet 152层，ImageNet top5错误率3.5%。

《*Deep Residual Learning for Image Recognition*》，CVPR 2016

深度学习方法推动视觉进展

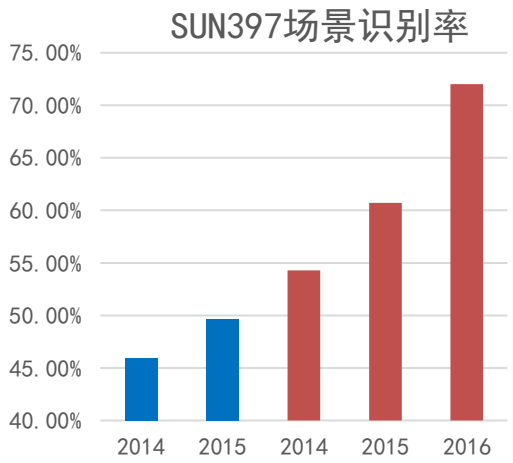
深度神经网络已被成功用于物体识别、场景分类、行为识别等视觉核心任务，极大地推动了计算机视觉的发展

视觉三大问题



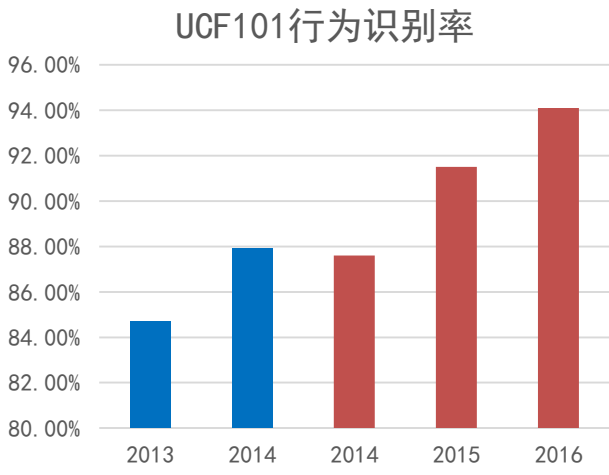
物体识别

回答“有什么”



场景分类

回答了“在哪里”



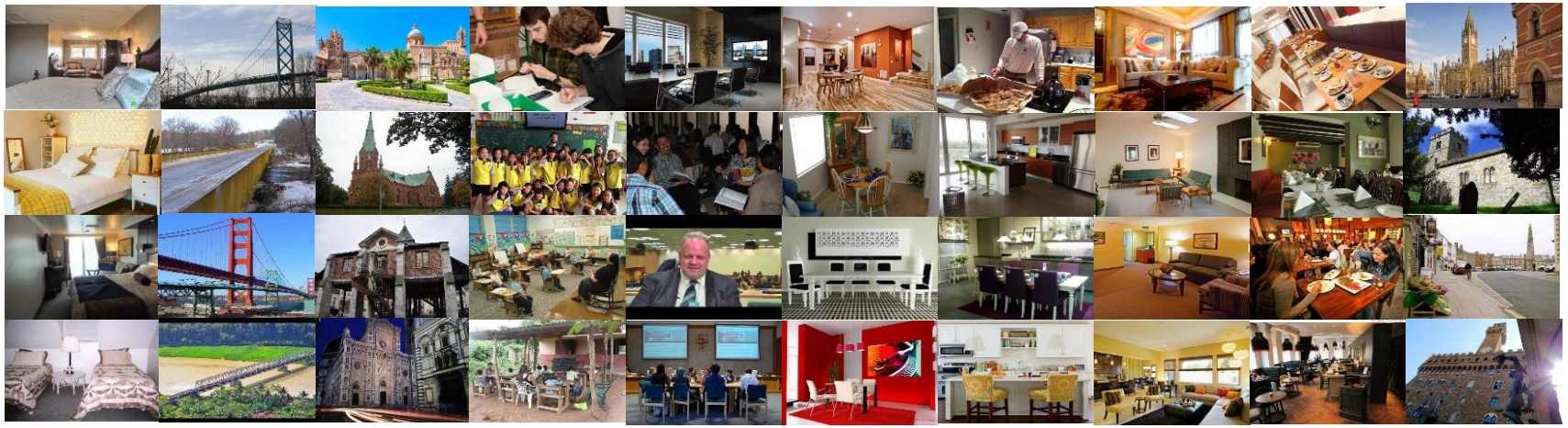
行为识别

回答“干什么”

Outline

- Overview
- Deep learning for large scale scene classification
  - Knowledge Guided Disambiguation for Scene Recognition with Multi-Resolution CNNs
- Deep learning approaches for action recognition
  - Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors (CVPR 2015)
  - Actioness estimation (CVPR 2016)
  - Real-time Action Recognition with Enhanced Motion Vector CNNs (CVPR 2016)
  - Deep Segmental Network (NO 1 in ActivityNet 2016)
- Conclusions

Problem of Scene Classification



Examples of ten scene categories.

Place2 Scene Dataset for ImageNet 2016-2016

- Places2 -2015 scene recognition challenge:
  - 401 scene categories, each class containing from 4,000 to 30,000 images.
  - 8.1M images for training, 20k images for validation and 381k images for testing coming.
  - Dataset size is much bigger than ImageNet.
- Places2 -2016 scene recognition challenge:
  - 365 scene categories, each class containing from 4,000 to 40,000 images.
  - 8M images for training, 36k images for validation and 328k images for testing coming.
- Scene recognition is challenging:
  - The concept of scene is more subjective and high level than object.
  - Larger intra-class variations (visual inconsistency).
  - Smaller inter-class variations (label ambiguity).

B. Zhou, A. Khosla, A. Lapedriza, A. Torralba and A. Oliva , *Places2: A Large-Scale Database for Scene Understanding*, in *Arxiv, 2015.*

Visual Inconsistency

Kitchen

Coffe Shop

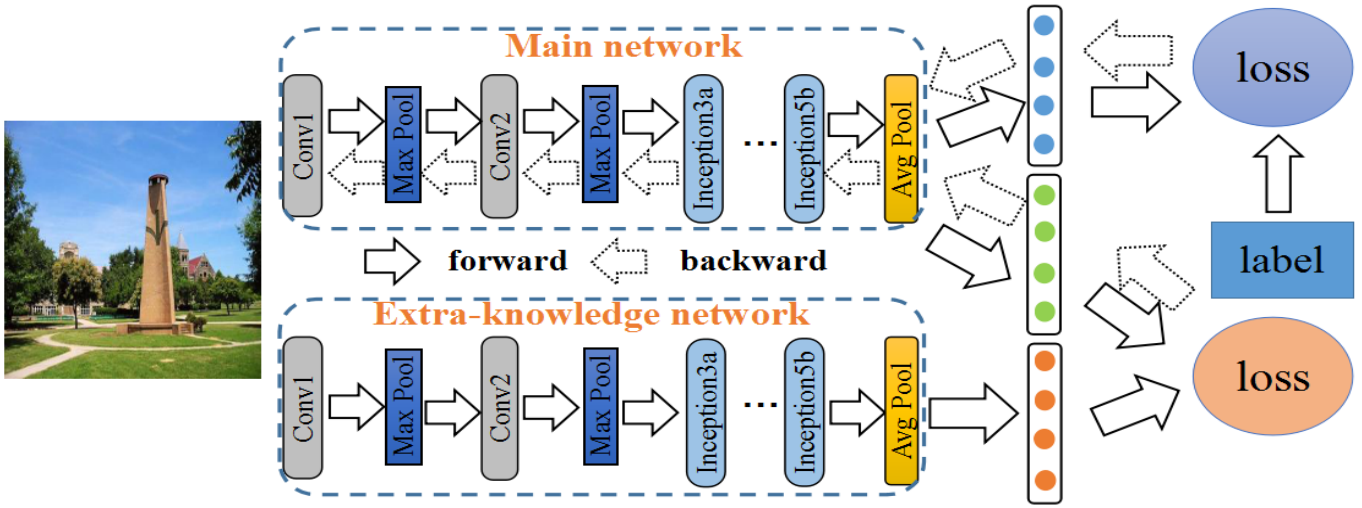
Label Ambiguity Example

cubicle office

office cubicles

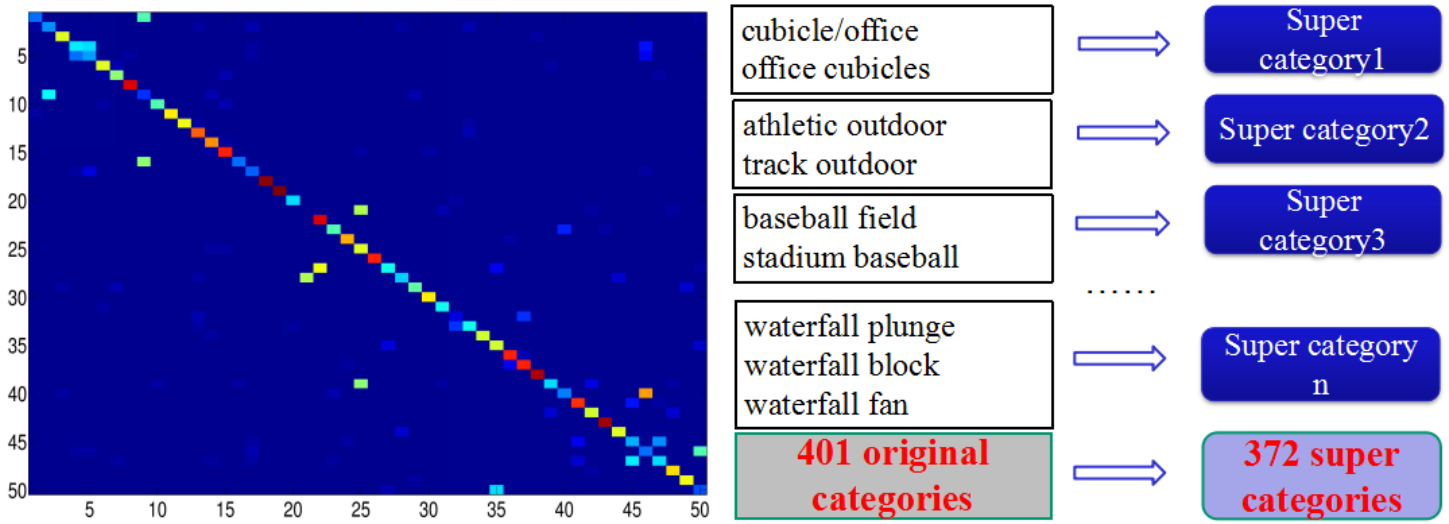
Knowledge Guided Disambiguation

Knowledge from networks trained on other datasets



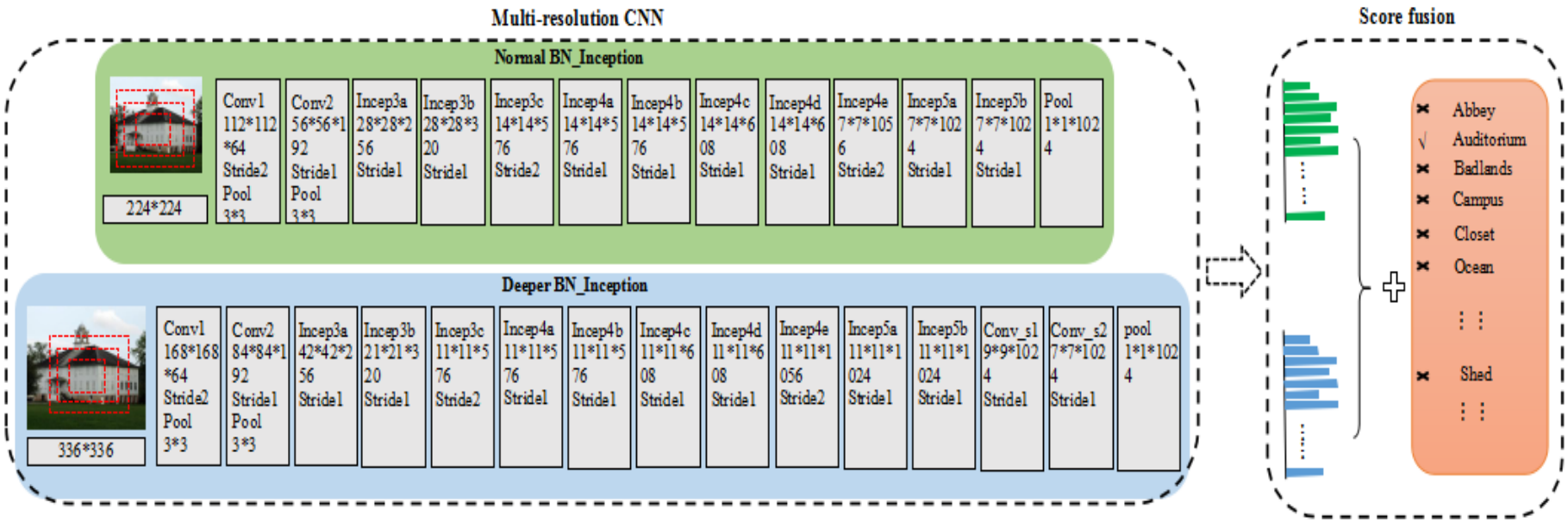
- In previous scenario, all the images belonging to the same super category are constrained to have the same label, without considering the difference between images.
- We propose to automatically assign a soft code to each image, which is able to better encode the visual information of natural images.
- In the soft code space, the images from easily confused categories are equipped with similar codes.
- Finally, we design a multi-task framework to predict both hard code and soft code

Knowledge from confusion matrix



- We propose a hierarchical strategy to merge similar categories into a super category, according to the confusion matrix on the validation data.
- The images of different scene categories, that belong to the same super category, will be given the same label.
- Totally, we reduce the number of scene categories from the Places2 dataset into 372 super categories.
- During test phase, we equally divide the score of super categories into their sub categories.

Multi-Resolution CNNs



Implementation details

- **Architectures:**
  - Low resolution: image (256*256), crop(224*224), inception2 network [2]
  - High resolution, image (384*384), crop(336*336), inception2+2 convs
- **Knowledge networks:**
  - Object nets: inception2 trained with ImageNet
  - Scene nets: inception2 trained with Places205
  - Currently, knowledge disambiguation only for low resolution CNNs
- **Implementation details:**
  - Resample images to balance the class distribution
  - Data augmentation: fixed crop, scale jittering, horizontal flipping [1,6]
  - Toolbox: we use a multi-GPU extension of Caffe, which is publicly available: https://github.com/yjxiong/caffe.git

L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, Towards Good Practices for Very Deep Two-Stream ConvNets, in *Arxiv, 2015.*

Experimental results

Method	Imagenet(top1/top5)	Places(top1/top5)	Places2(top1/top5)
AlexNet	40.7%/18.2%	50.0%/-	57.0%/-
VGGNet	27.0%/8.8%	39.4%/11.5%	52.4%/-
Normal BN-Inception	24.7%/7.2%	38.1%/11.3%	48.8%/17.4%
Deeper BN-Inception	23.7%/6.6%	37.8%/10.7%	48.0%/16.7%
Multi-resolution CNN	**21.8%/6.0%**	**36.4%/10.4%**	**47.4%/16.3%**

Performance of Multi-Resolution CNNs on the validation data from the datasets of ImageNet, Places and Places2.

Experimental Results

Model	MIT Indoor67	SUN397
ImageNet-VGGNet-16 [31]	67.7%	51.7%
Places205-AlexNet [1]	68.2%	54.3%
Places205-GoogLeNet [46]	74.0%	58.8%
DAG-VggNet19 [8]	77.5%	56.2%
Places205-CNDS-8 [47]	76.1%	60.7%
Ms-DSP [48]	78.3%	59.8%
Places205-VGGNet-16 [49]	81.2%	66.9%
LS-DHM [46]	83.8%	67.6%
Multiple Models [50]	86.0%	70.7%
Three [51]	86.0%	70.2%
Places2-Deeper-BN-Inception	**86.7%**	**72.0%**

Performance of our pretrain models with Multi-Resolution CCNs on MIT Indoor67 and SUN397.

Experimental Results

Rank	Team	Top1
1	WM	16.9%
2	SIAT_MMLAB(our)	17.4%
3	Qualcomm	17.6%
4	Trimps-Soushen	18.0%
5	NTU_Rose	19.3%

Imagenet2015

Rank	Team	Top1
1	SIAT_MMLAB(our)	91.6%
2	SJTU-ReadSense	90.4%
3	TEG Rangers	88.7%
4	ds-cube	83.0%
	Google (last year winner)	91.2%

LSUN2016

Conclusions on scene classification

- Large scale scene datasets with many categories come along with increased ambiguity between the class labels (e.g. baseball field vs. stadium baseball).
  - Knowledge guided disambiguation aims to regularize CNN training with extra knowledge and improve the generalization capacity.

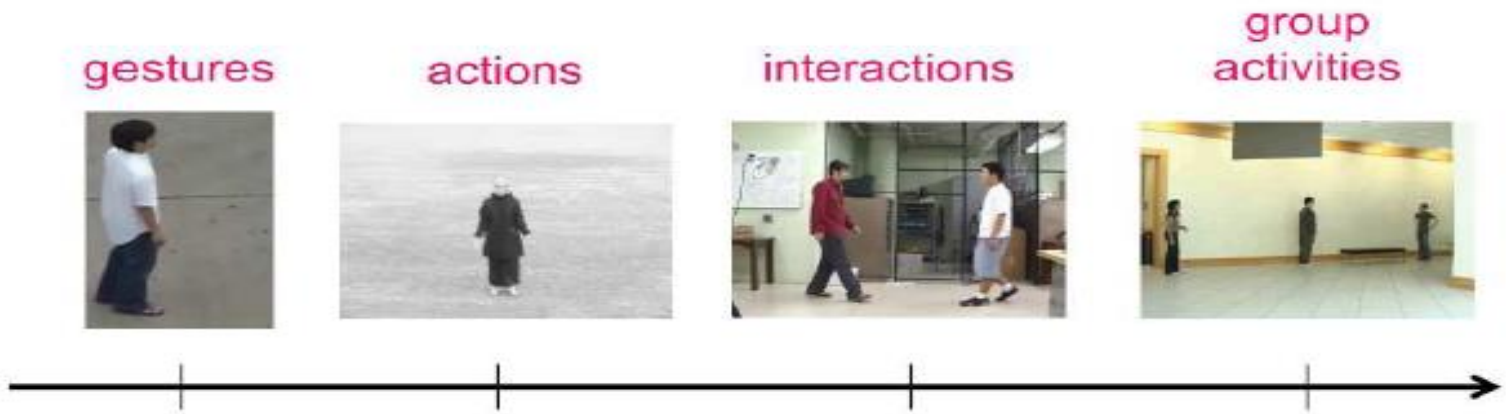
- Scene or Places, defined by containing objects, spatial layout, human events, and global contexts, are more high-level concepts.
  - Multi-Resolution CNNs take images of different sizes as input and capture visual information from different levels.

Outline

- Overview
- Deep learning for large scale scene classification
  - Knowledge Guided Disambiguation for Scene Recognition with Multi-Resolution CNNs
- Deep learning approaches for action recognition
  - Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors (CVPR 2015)
  - Actioness estimation (CVPR 2016)
  - Real-time Action Recognition with Enhanced Motion Vector CNNs (CVPR 2016)
  - Deep Segmental Network (NO 1 in ActivityNet 2016)
- Conclusions

What is action?

Definition: Action is what an agent can do.



Four levels of human activities:

- **Gesture**: elementary poses and movements of a person's body part.
- **Action**: single person activity that may be composed of a sequence of multiple gestures.
- **Interaction**: activity that involves two or more persons and object.
- **Group activity**: group activity performed by conceptual groups composed of multiple persons and objects

Human Action Understanding

- The goal of human action recognition is to automatically detect and classify ongoing activities from an input video (i.e. a sequence of images frames).
  - Human vision system is very effective in perceiving and predicting actions through visual information.
  - A basic problem in computer vision, with wide applications.

Action recognition

Walk

Run

Boxing

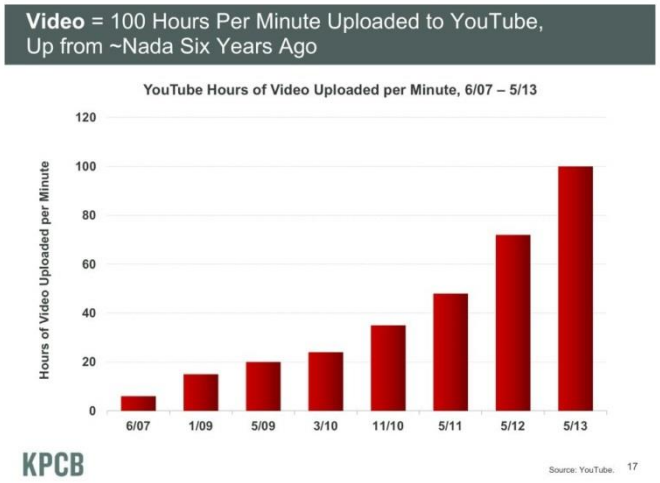
…

Application 1-Surveillance

- Ubiquitous surveillance cameras in public places.
  - More than 60M in China
  - A person is monitored average 300 times / day in London.
  - Current surveillance system mainly record without understanding human action and event in video.
- Understanding human activity is important for intelligent surveillance system.

Application 2- Online Videos

- Explosive growth of online videos



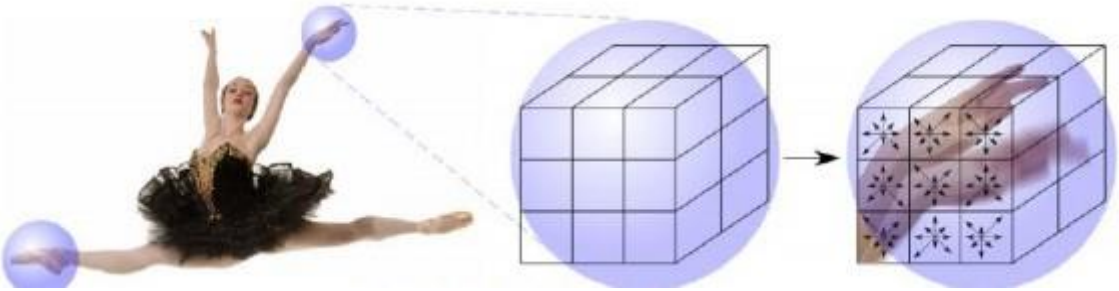
YOUKU 优酷

Users > 263million
User online time>22billion hrs/season

- Many online videos contain action or activity
- Action recognition has important applications in video tagging and content based video retrieval.

Representation of Action Videos

- Like many CV tasks, representation of videos is key in action recognition.
- Local features provided a popular and effective representation for action video.
  - Detectors of interest points/ trajectories
  - Motion and Appearance Descriptors of 3d cuboids

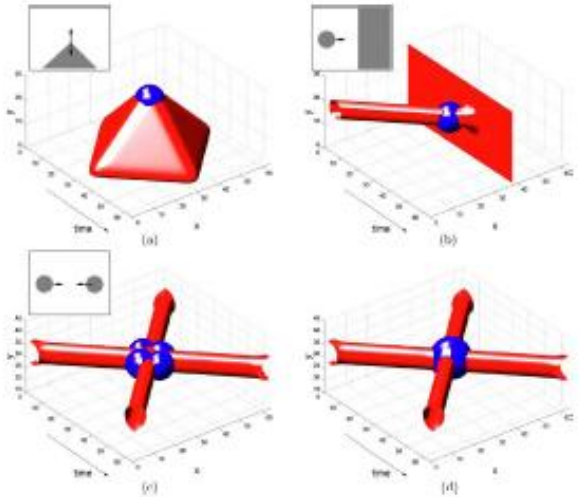


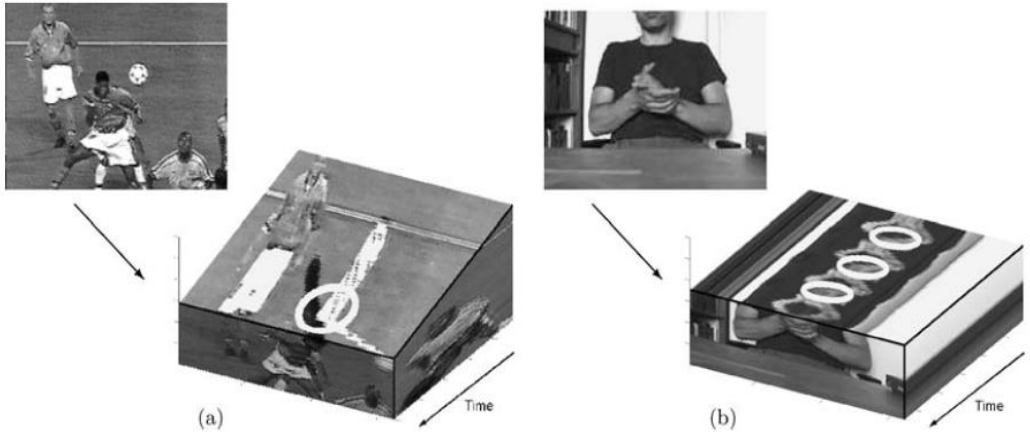
Detectors Descriptors

Detectors of interest points

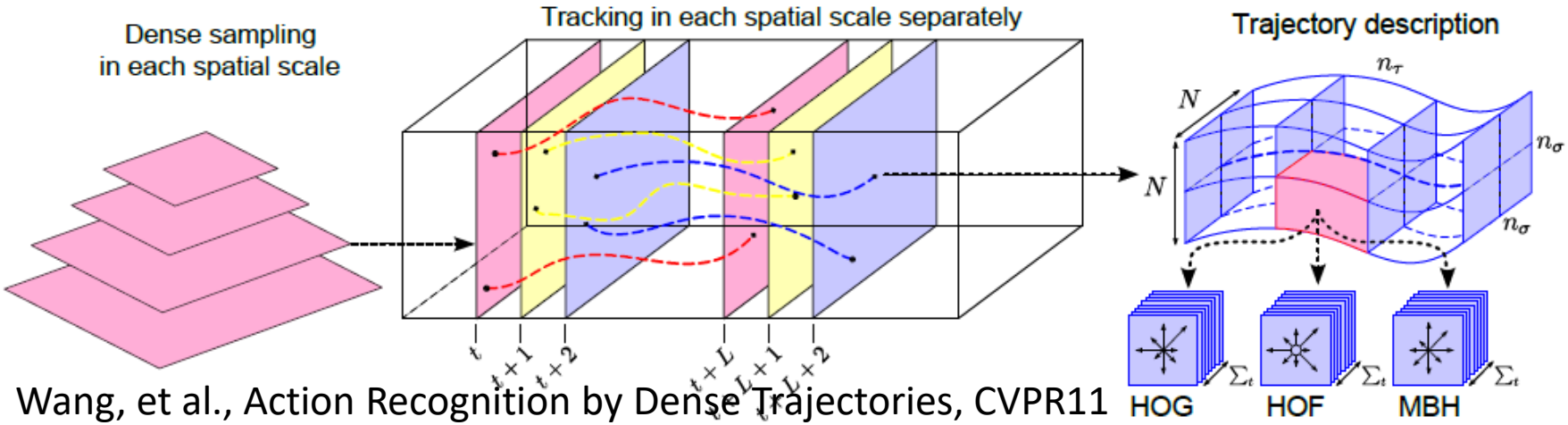
- Spatial-temporal interest points

I.Laptev, "On Space-Time Interest Points", IJCV 2005.



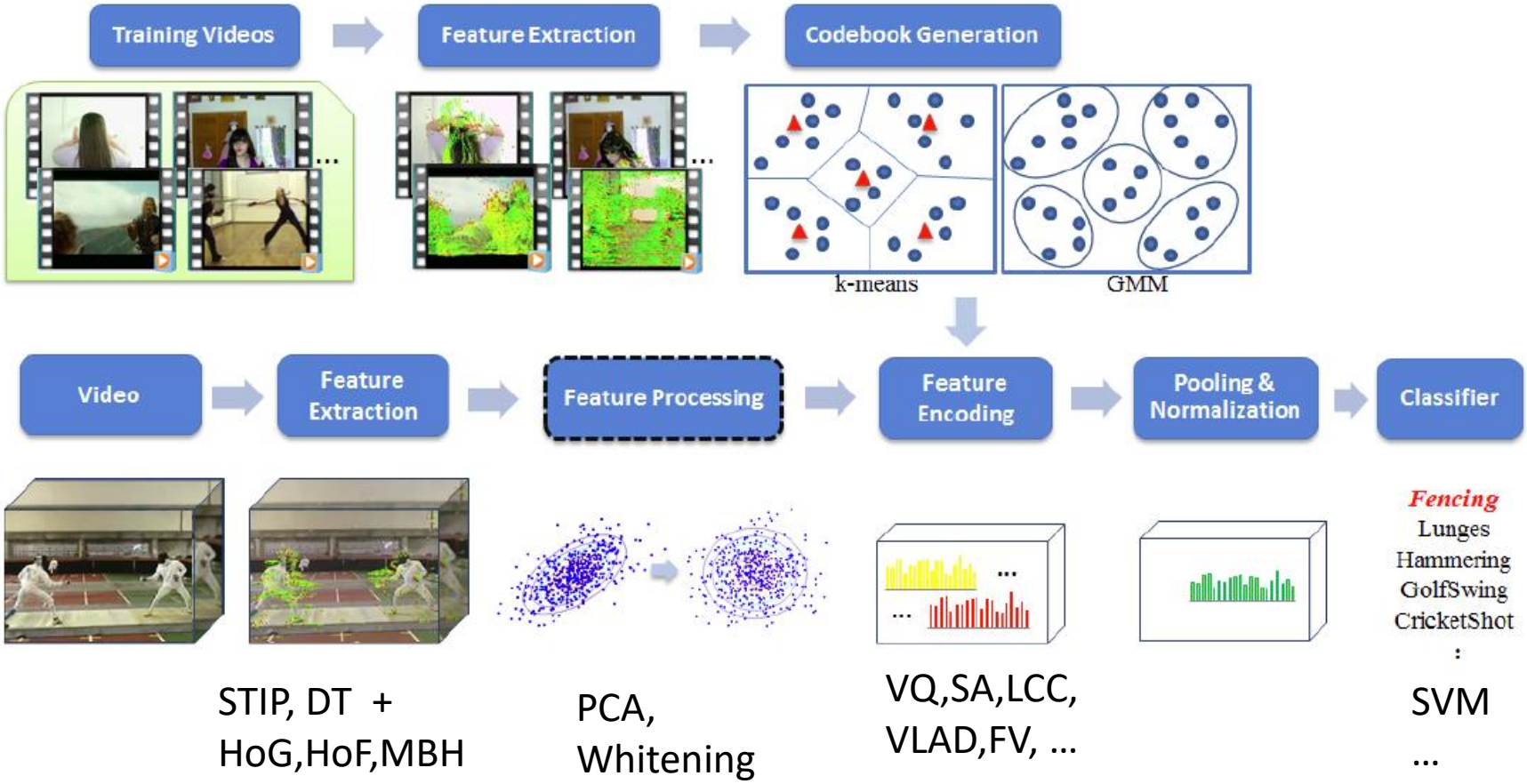


- Dense trajectories



Wang, et al., Action Recognition by Dense Trajectories, CVPR11

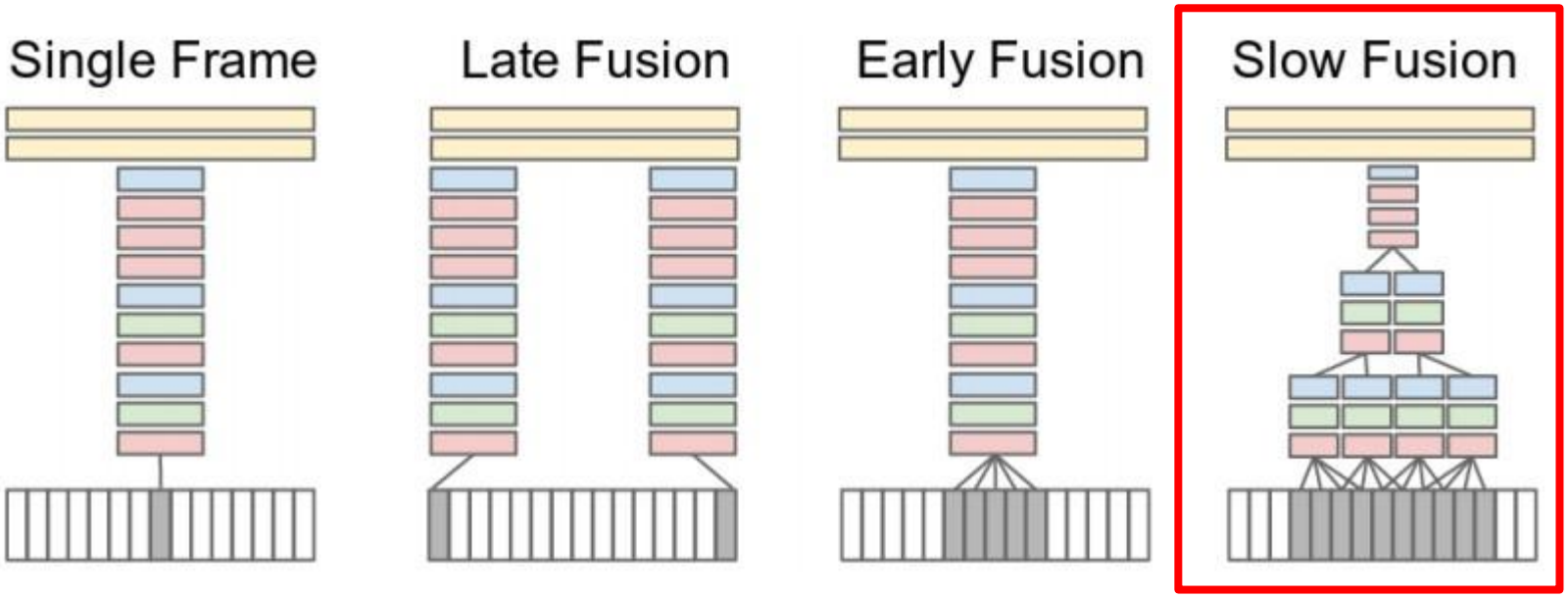
A typical BoVW pipeline for action recognition



Xiaojiang Peng, Limin Wang, Xingxing Wang, Yu Qiao, "Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice" CVIU, 2016

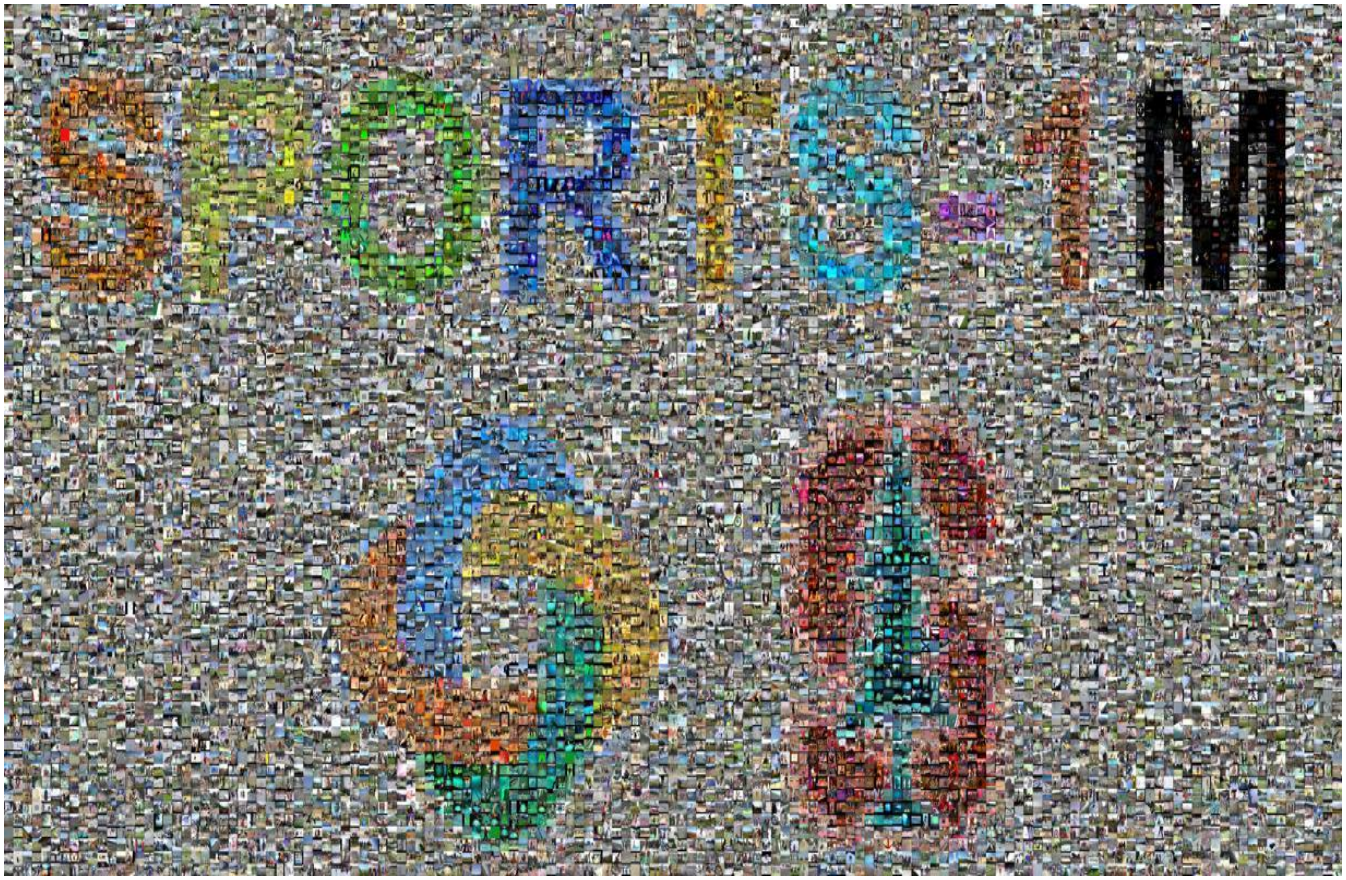
Spatio-Temporal ConvNets

spatio-temporal convolutions;
worked best.



[Large-scale Video Classification with Convolutional Neural Networks, Karpathy et al., 2014]

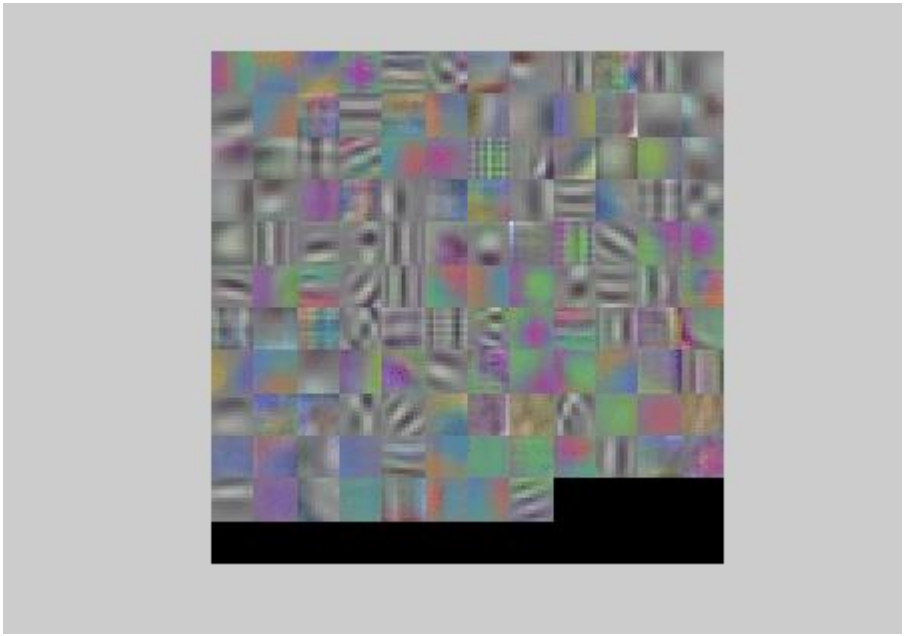
Sports-1M Dataset



1 million videos
487 sports classes

[Large-scale Video Classification with Convolutional Neural Networks, Karpathy et al., 2014]

Spatio-Temporal ConvNets



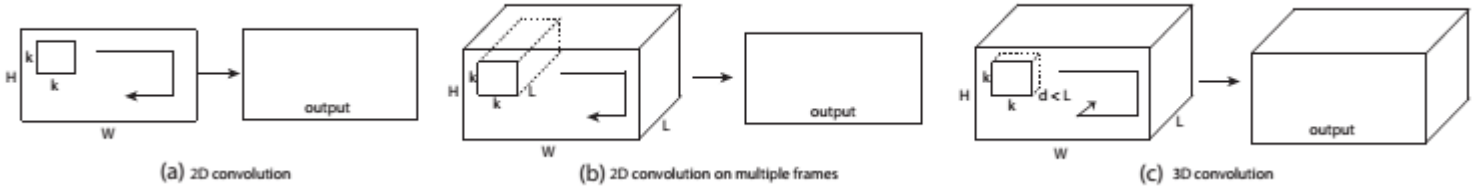
三维滤波器

Model	3-fold Accuracy
Soomro et al [22]	43.9%
Feature Histograms + Neural Net	59.0%
Train from scratch	41.3%
Fine-tune top layer	64.1%
Fine-tune top 3 layers	**65.4%**
Fine-tune all layers	62.2%

识别率一般

Ours :87.1 % with iDT+FV

C3D



2D Convolution → 3D Convolution

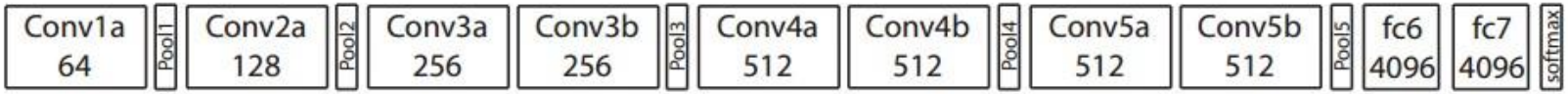


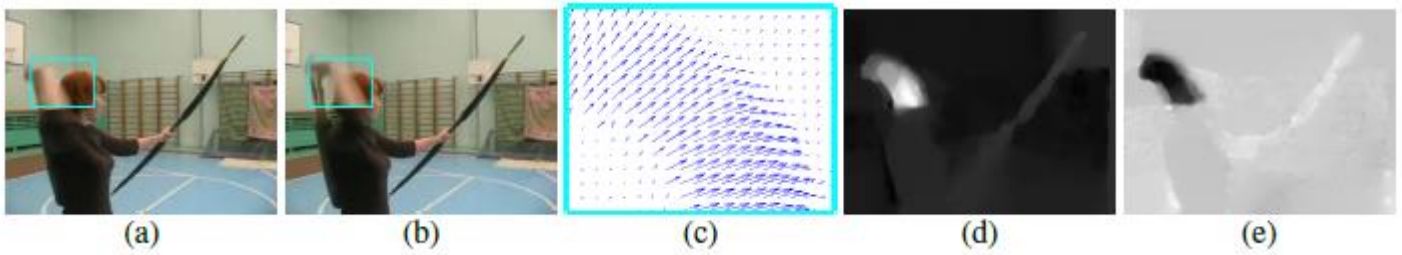
Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

C3D: 3D VGGNet

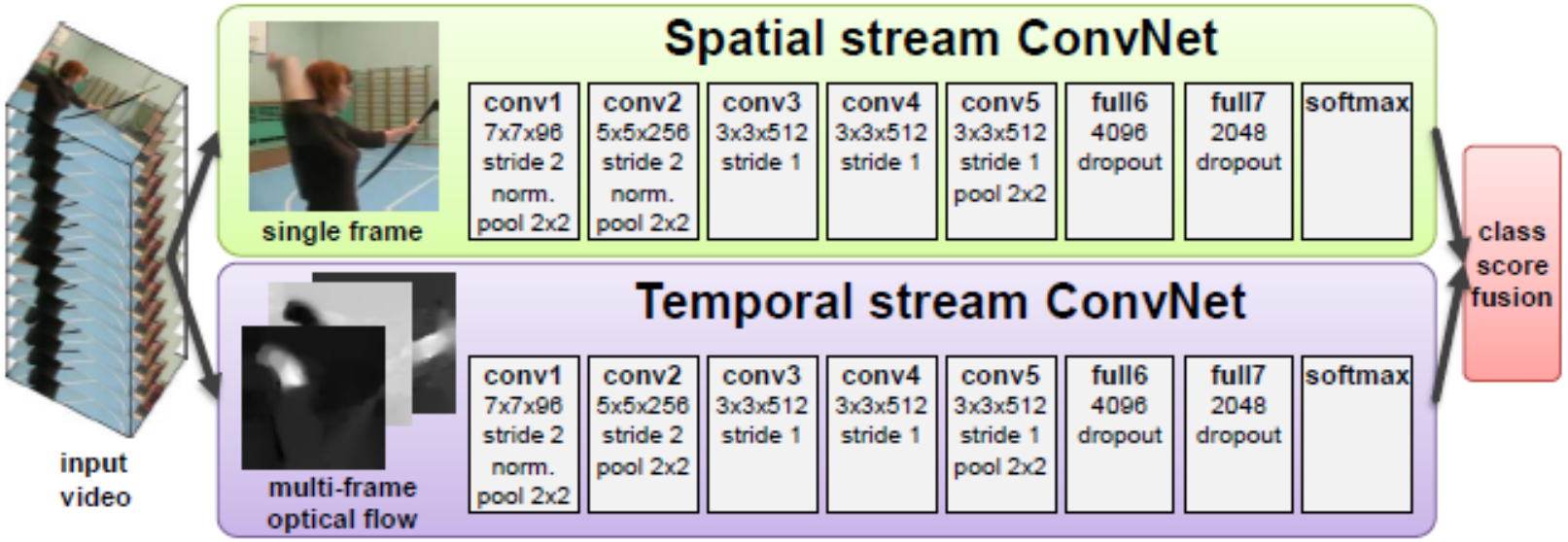
[Learning Spatiotemporal Features with 3D Convolutional Networks, Tran et al. 2015]

Two stream CNN for action recognition

Treat optical flow as images



Train spatial CNN from images and temporal CNN from optical flows



Karen Simonyan Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos“, NIPS, 2014

Recognition performance of Two-steam

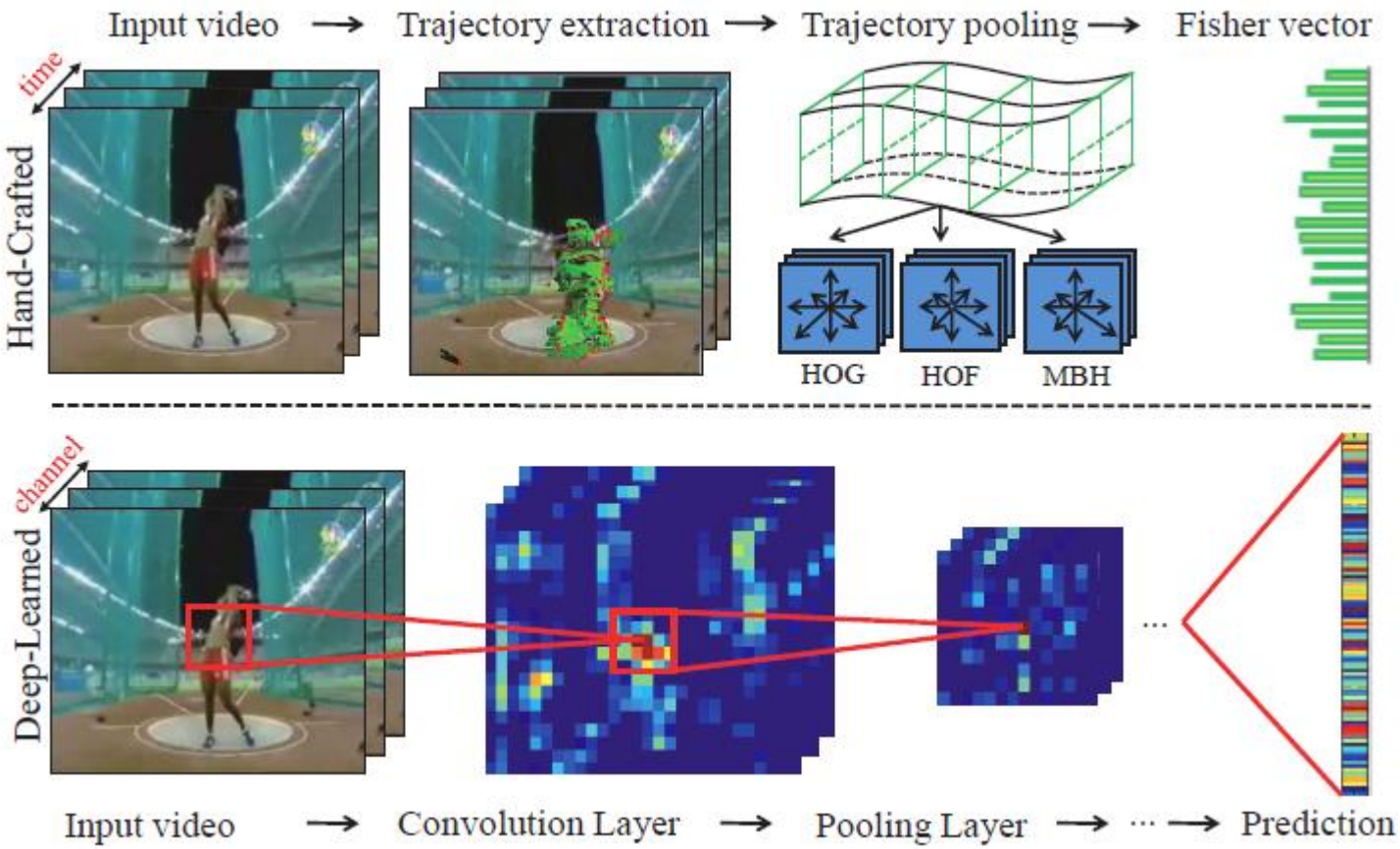
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

Two-stream version works much better than either alone.

[Two-Stream Convolutional Networks for Action Recognition in Videos, **Simonyan** and Zisserman 2014]

[T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," 2011]

Trajectory-Pooled Deep-Convolutional Descriptors



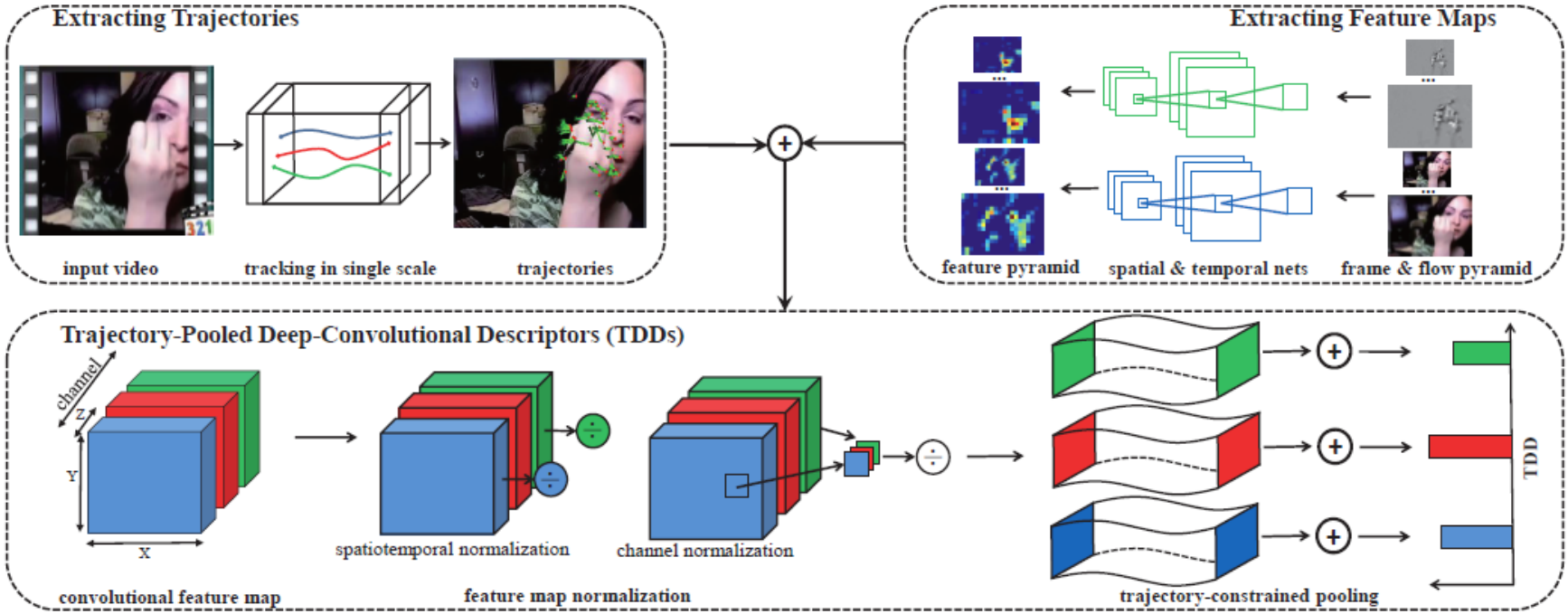
How to combine the merits of two approaches

Limin Wang, Yu Qiao, Xiaoou Tang "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015

Framework of the proposed methods

Propose *trajectory-pooled deep convolutional descriptor* (TDD) to integrate the key factors from handcrafted and deep approaches.

- Utlize two-stream ConvNets to obtain multi-scale deep convolutional features.
- Pool the local ConvNet responses over the spatiotemporal tubes centered at the trajectories.



Train ConvNet Architectures

Layer	conv1	pool1	conv2	pool2	conv3	conv4	conv5	pool5	full6	full7	full8
size	7×7	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
stride	2	2	2	2	1	1	1	2	-	-	-
channel	96	96	256	256	512	512	512	512	4096	2048	101
map size ratio	1/2	1/4	1/8	1/16	1/16	1/16	1/16	1/32	-	-	-
receptive field	7×7	11×11	27×27	43×43	75×75	107×107	139×139	171×171	-	-	-

Obtain a set of convolutional feature maps from input video

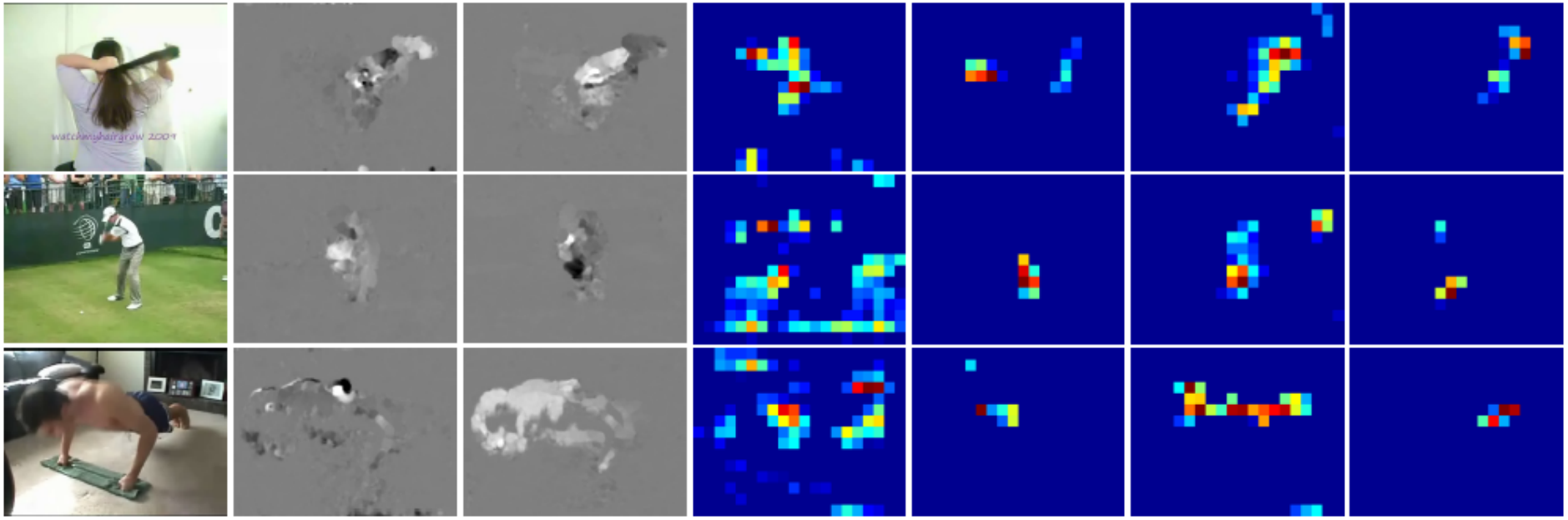
$$\mathbb{C}(V) = \{C_1^s, C_2^s, \cdots, C_M^s, C_1^t, C_2^t, \cdots, C_M^t\},$$

feature map of spatial net feature map of temporal net

Spatiotemporal Normalization: $\widetilde{C}_{st}(x,y,z,n) = C(x,y,z,n)/\mathrm{maxV}_{st}^{n},$

Channel Normalization $\widetilde{C}_{ch}(x,y,z,n) = C(x,y,z,n)/\mathrm{maxV}_{ch}^{x,y,z}$

Example of Convolutional Features



(a) RGB (b) Flow-x (c) Flow-y (d) S-conv4 (e) S-conv5 (f) T-conv4 (g) T-conv5

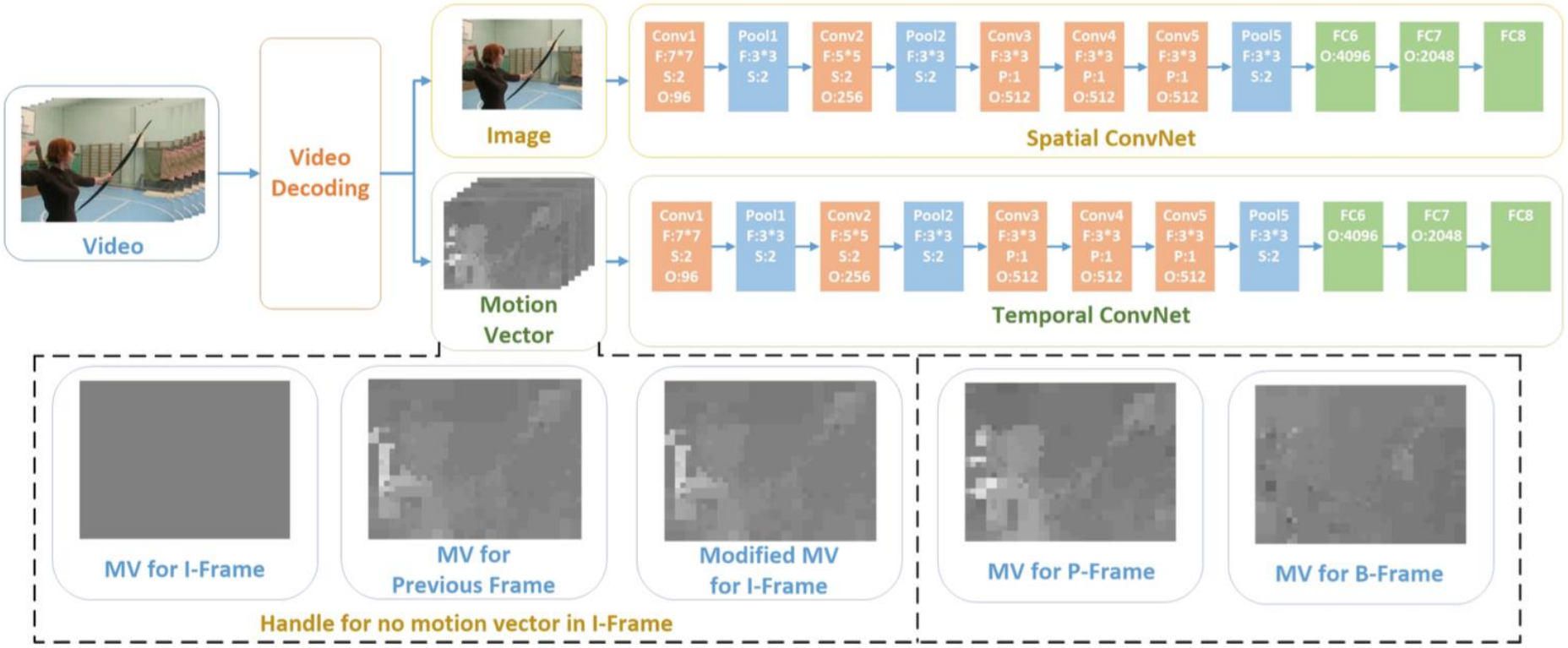
Convolutional feature maps are relatively sparse and exhibit high correlation with the action areas.

Performance evaluation

Algorithm	HMDB51	UCF101
HOG [31, 32]	40.2%	72.4%
HOF [31, 32]	48.9%	76.0%
MBH [31, 32]	52.1%	80.8%
HOF+MBH [31, 32]	54.7%	82.2%
iDT [31, 32]	**57.2%**	**84.7%**
Spatial net [24]	40.5%	73.0%
Temporal net [24]	54.6%	83.7%
Two-stream ConvNets [24]	**59.4%**	**88.0%**
Spatial conv4	48.5%	81.9%
Spatial conv5	47.2%	80.9%
Spatial conv4 and conv5	**50.0%**	**82.8%**
Temporal conv3	54.5%	81.7%
Temporal conv4	51.2%	80.1%
Temporal conv3 and conv4	**54.9%**	**82.2%**
TDD	63.2%	90.3%
TDD and iDT	**65.9%**	**91.5%**

Motion Vector CNN

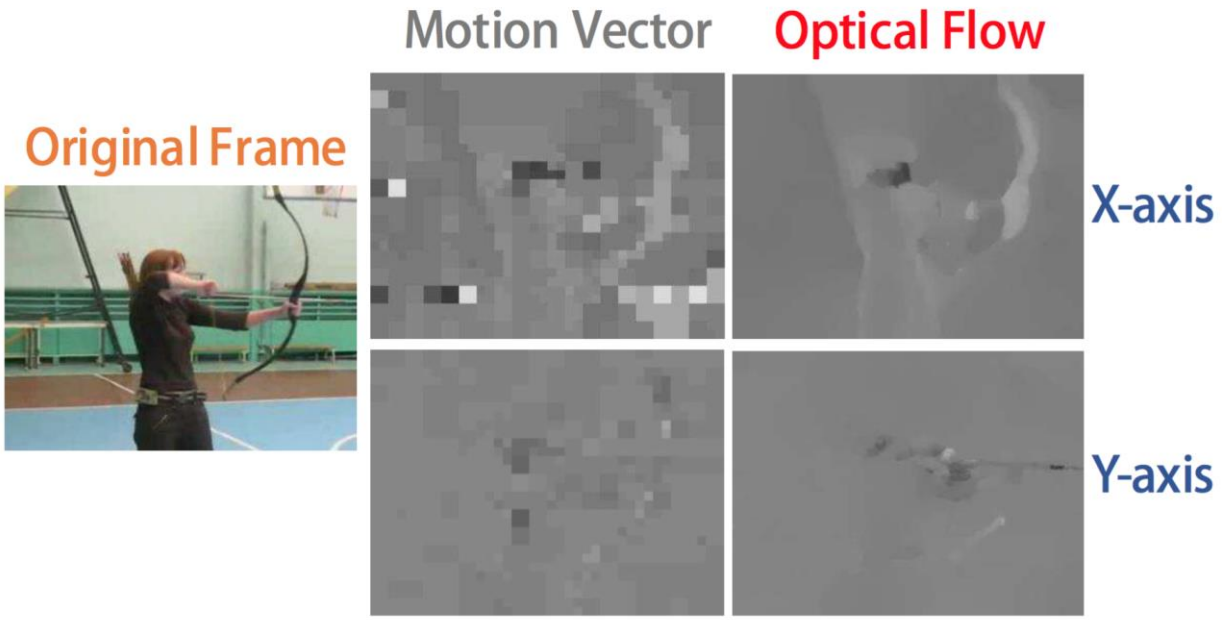
- Many deep learning approaches for video based action recognition are computationally expensive, due to the calculation of optical flows
- Motion vector also includes motion information of local regions



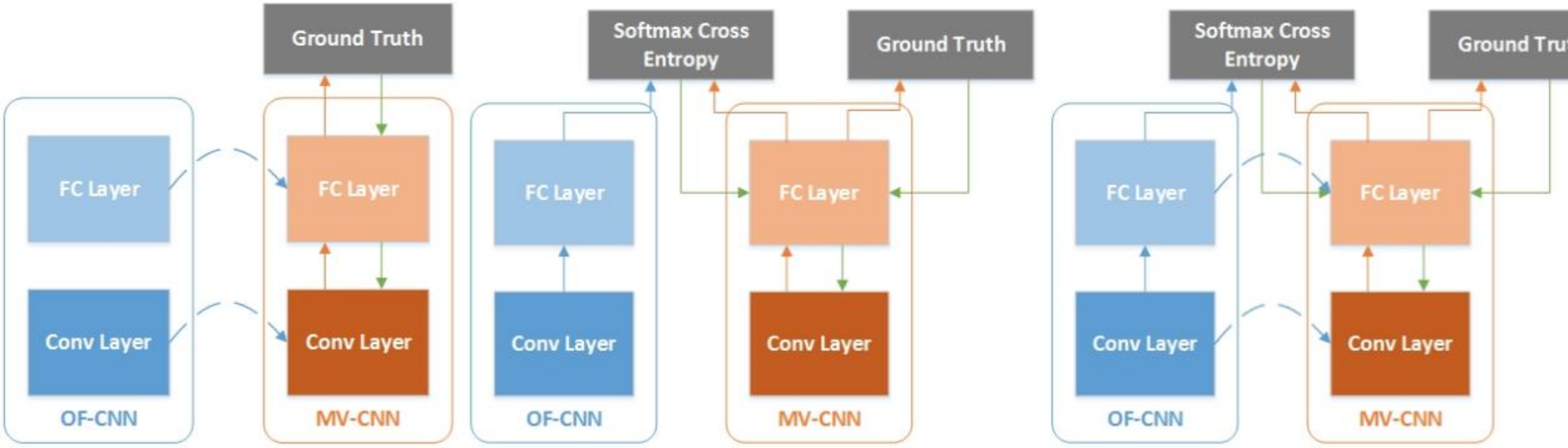
B. Zhang, L. Wang, Z. Wang, Yu Qiao, and H. Wang " Real-time Action Recognition with Enhanced Motion Vector CNNs ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016

Motion Vector

- MV only contains coarse motion information and lack fine grained details.
- MV includes much more noise than optical flow
- How to train effective CNN with MV as input?



Enhanced Motion Vector CNN



(a) Strategy 1: Teacher Initialization

Fine tune Optical Flow CNN

(b) Strategy 2: Supervision Transfer

OF-CNN output as supervision

(c) Strategy 3: Combination

Strategy 1+2

Samples of filters for Conv1 layer. Left to right: MV- CNN, EMV-CNN and OF-CNN.

Experiments

- Performance comparison of three teaching strategies

Temporal CNN	Accuracy
OF-CNN [23]	81.2%
MV-CNN trained from scratch	74.4%
EMV-CNN with ST	77.5%
EMV-CNN with TI	78.2%
EMV-CNN with ST+TI	**79.3%**

- Speed and Accuracy comparison of 5 crops+mirror with 1 crop

EMV+RGB-CNN	UCF101(Split1) (fps)	THUMOS14 (fps)
5 crops+mirror	88.0	89.0
1 crop	390.7	403.2

EMV+RGB-CNN	UCF101(Split1)	THUMOS14
5 crops+mirror	86.6%	61.2%
1 crop	85.7%	61.5%

Comparison with state-of-the-art

Speed and Accuracy Performance of UCF-101 and THUMOS 14

	Accuracy	FPS
MV+FV (CPU) (re-implement) [12]	78.5%	132.8
C3D (1 net) (GPU) [27]	82.3%	313.9
C3D (3 net) (GPU) [27]	85.2%	-
iDT+FV (CPU) [28]	85.9%	2.1
Two-stream CNNs (GPU) [23]	88.0%	14.3
EMV+RGB-CNN	**86.4%**	**390.7**

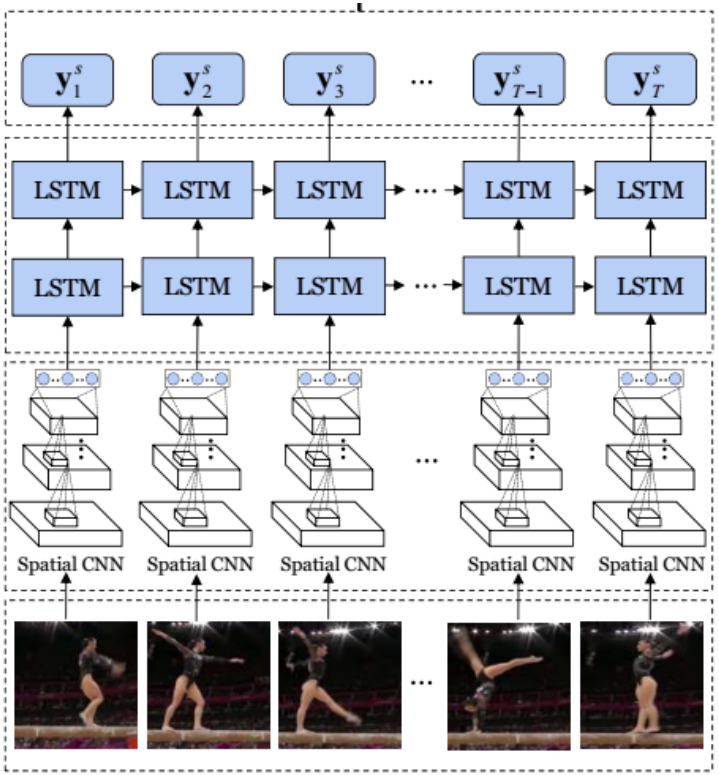
UCF-101

	Accuracy	FPS
Objects (GPU) [8]	44.7%	-
iDT+CNN (CPU+GPU) [32]	62.0%	< 2.38
Motion (iDT+FV) (CPU) [8]	63.1%	2.38
Objects+Motion (CPU+GPU) [8]	71.6%	< 2.38
EMV+RGB-CNN	**61.5%**	**403.2**

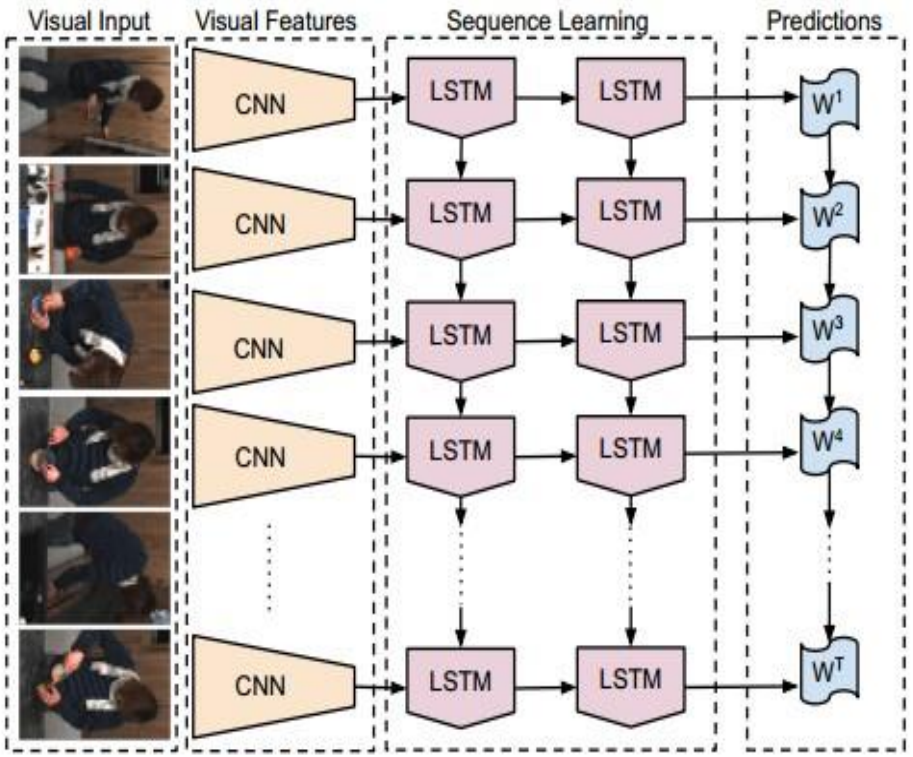
THUMOS 14

Speed: 400 frames/s with GPU

Recurrent Neural Network/LSTM for Action Recognition



Wu Z, Wang X, Jiang Y, et al. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification[C]. acm multimedia, 2015: 461-470.



[Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al., 2015]

ChaLearn Looking at People Challenge

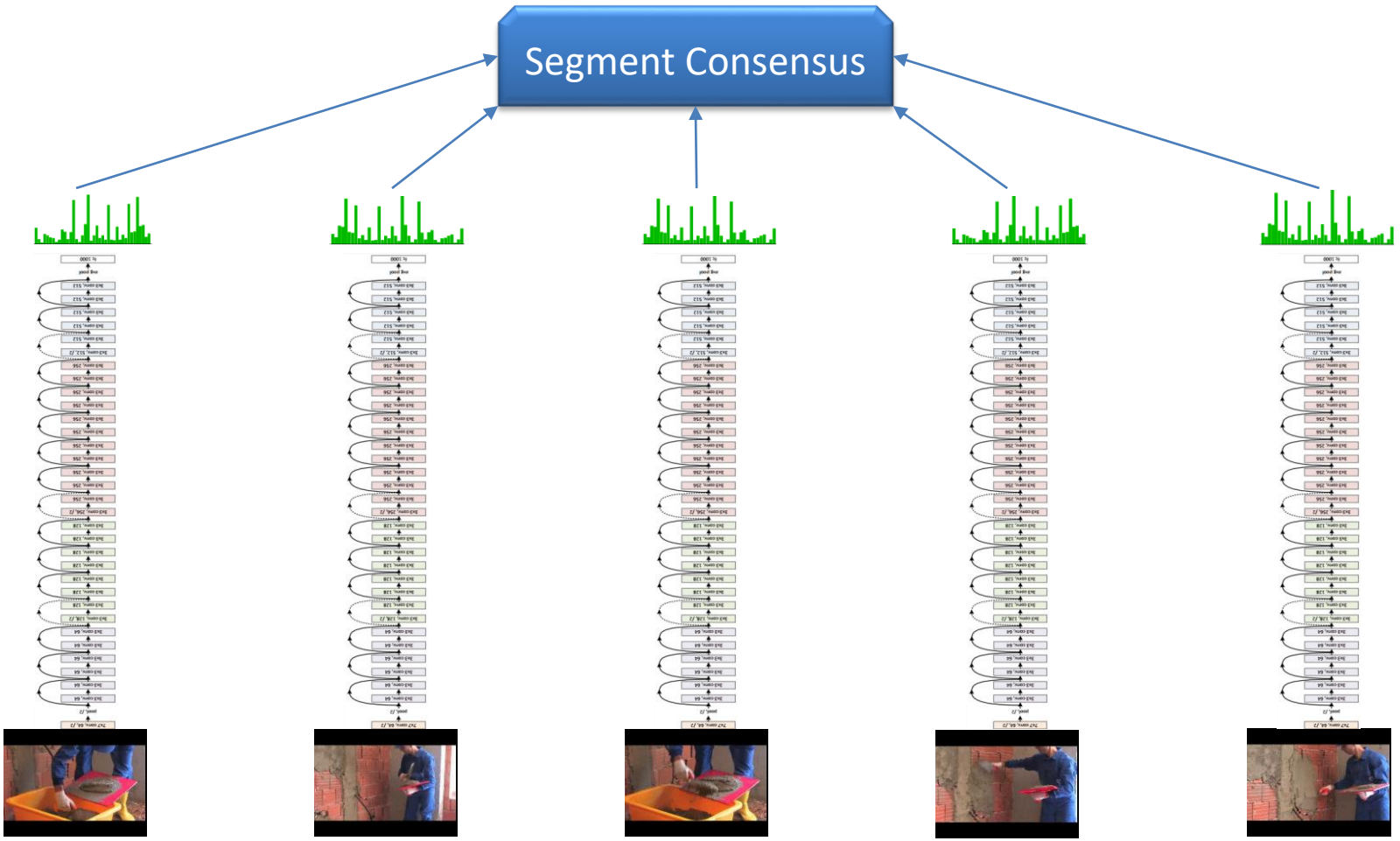
- ChaLearn 2014 : the 1st on both track 1 and track2, and 4th for track3
- ChaLearn 2015 : 1st winner of both tracks for action recognition and cultural event recognition

(a) Track 2: Action/Interaction Recognition

(b) Track 3: Gesture Recognition

X.J. Peng, L.M. Wang, Z.W. Cai, and Y. Qiao Action and Gesture Temporal Spotting with Super Vector Representation in Proceedings of European Conference on Computer Vision Workshop (ECCVW), Zurich, Switzerland, 2014.

Deep Segmental Network



From Y. Xiong 2016

ActivityNet Challenge in CVPR 2016

ACTIVITYNET

200 categories， 648 hrs video， 10k for training , 5k for testing

Achieve NO 1 in classification task in ActivityNet 2016 among 24 teams.

Validation Set	mAp	Top-3 Acc.
Visual	90.4%	95.2%
Audio	15.2%	29.1%
Visual + Audio	90.9%	95.6%
Testing Set	**mAP**	**Top-3 Acc.**
Visual CNN (Single)	91.2%	95.6%
Final Ensemble	93.2%	96.4%

http://activity-net.org/challenges/2016/

Publications

1. Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," Proc. European Conference Computer Vision (ECCV), 2016
2. Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang,Luc Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," Proc. European Conference Computer Vision (ECCV), 2016
3. Zhi Tian, Weilin Huang, Tong He, Pan He, Yu Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network," Proc. European Conference Computer Vision (ECCV), 2016
4. L. Wang, Yu Qiao, X. Tang, and L. Van Gool " Actionness Estimation Using Hybrid Fully Convolutional Networks," Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016
5. B. Zhang, L. Wang, Z. Wang, Yu Qiao, and H. Wang " Real-time Action Recognition with Enhanced Motion Vector CNNs," Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016
6. W. Zhu, J. Hu, G. Sun, X. Cao, Yu Qiao "A Key Volume Mining Deep Framework for Action Recognition," Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016
7. Limin Wang, Y. Qiao, Xiaoou Tang, " Motionlets: Mid-Level 3D Parts for Human Motion Recognition," Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015
8. L. Wang, *Yu Qiao*, X. Tang, " Video Action Detection with Relational Dynamic-Poselets ," Proc. European Conference Computer Vision (ECCV), 2014
9. X. Peng ,L. Wang, *Yu Qiao*, Q. Peng, " Accelerated High-Order Super Vectors for Visual Recognition ," Proc. European Conference Computer Vision (ECCV), 2014
10. X. Peng, *Yu Qiao*, Q. Peng, " Action Recognition with Stacked Fisher Vectors ," Proc. European Conference Computer Vision (ECCV), 2014
11. W. Huang, *Yu Qiao*, X. Tang, " Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees ," Proc. European Conference Computer Vision (ECCV), 2014
12. L. Wang, *Yu Qiao* , and X. Tang , "Latent Hierarchical Model of Temporal Structure for Complex Activity Classification ", IEEE Transactions on Image Processing (TIP), Vol. 23, No. 2, 2014.
13. Z. Cai and Y. Qiao " Multi-View Super Vector for Action Recognition ," Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2014 **Oral with 5.75% acceptance rate**
14. L. Wang, Y. Qiao, and X. Tang "Mining Motion Atoms and Phrases for Complex Action Recognition," International Conference on Computer Vision (ICCV), 2013
15. Limin Wang, Y. Qiao, Xiaoou Tang, " Motionlets: Mid-Level 3D Parts for Human Motion Recognition," Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2013
16. X.Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition", Proc. BMVC, 2013
17. Xingxing Wang, Limin Wang, Y. Qiao," A Comparative Study of Encoding, Pooling and Normalization Methods for Action Recognition," Proc. Asian Conference on Computer Vision (ACCV), 2012.
18. X.J. Peng, *Yu Qiao* , Q. Peng , Motion Boundary Based Sampling and 3D Co-occurrence Descriptors for Action Recognition", Image and Vision Computing, vol.32, no 9, pp. 616-C628, 2014

Thank you!

Q&A