关于在静态彩色图像上的动作识别调研情况

By 紫梦 lan (2015.05.08)

数据集(Dataset)

目前在单张静态彩色图上做活动识别/分类的数据集不多，多数论文所用的数据集有以下几个:

1 Pascal Voc Action Dataset 系列，如 Pascal Voc 2007, Pascal Voc 2010, Pascal Voc 2011, Pascal Voc 2012. 用的比较多的是 2012 和 2010 这两个数据集.

2 Willowactions Dataset

3 Stanford40 Dataset

4 PPMI Dataset

5 SUN Action Dataset

6 MPII pose Dataset

7 Six-Class Sports Dataset

下面就概要说下每个数据集的情况:

1 Pascal Voc 2012 Dataset, Pascal Voc 2011 Dataset, Pascal Voc 2010 Dataset, Pascal Voc 2007 Dataset (这里主要介绍 Pascal Voc 2012 Action Dataset)

该数据集有 9157 张静态彩色图像，其中 4588 张为训练集，4569 张为测试集，由 10 个动作类别组成:

Jumping
Phoning
Playing Instrument
Reading
Taking Photo
Riding Bike
Riding Horse
Runing
Using Computer
Walking

2 Willow Actions Dataset

该数据集有 911 张静态彩色图像，由 7 个动作类别组成:

Interacting with computer
Photographing
Playing music instrument
Riding bike
Riding horse
Running
Walking

对应的论文为:

V. Delaitre, I. Laptev and J. Sivic Recognizing human actions in still images: a study of

bag-of-features and part-based representations Proceedings of the 21st British Machine Vision Conference, Aberystwyth, September 2010, poster.

3 Stanford40 Dataset

该数据集有 9532 张静态彩色图像, 由 40 动作类别组成. 为每张静态彩色图像提供了一个执行动作的 subject 的 bounding box 和对应的活动类别. 每类动作大概有 180 到 300 左右的静态彩色图像.

对应的论文为:

B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.

4 PPMI Dataset

该数据集包含了人交互的 12 种乐器的数据集, 十二种乐器分别为:

Bassoon, Cello, Clarinet, Erhu,
Flute, French Horn, Guitar, Harp,
Recorder, Saxophone, Trumpet, Violin

对应的论文为:

Bangpeng Yao and Li Fei-Fei. Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

5 SUN Action Dataset

该数据集主要是场景的数据集, 也就是给一个场景的静态图像, 估计该场景会发生什么动作. 例如给一张羽毛球场的静态图片(里面可能没有 subject), 那么该图像很可能会发生打羽毛球的动作.

该数据集由 397 个场景类别中的 61 动作类别组成, 其中有 194 个户外场景的 38 个类别, 203 个户内场景的 23 个动作类别组成.

对应的论文为:

VU, T.H. and Olsson, C. and Laptev, I. and Oliva, A. and Sivic, J. Predicting Actions from Static Scenes. In ECCV, 2014.

6 MPII Pose Dataset

该数据集主要是 pose 的数据集, 但是它里面提供了 activity 的 label, 所以这里就把它当作动作的数据集. 该数据集大概有 25k 的静态彩色图像, 覆盖了大约 40k 的 subjects, 由 410 种 aactivities 构成.

对应的论文:

Mykhaylo Andriluka and Leonid Pishchulin and Peter Gehler and Schiele, Bernt. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In CVPR, 2014.

7 Six-Class Sports Dataset

对应的论文:

Abhinav Gupta, Aniruddha Kembhavi and Larry S. Davis, Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition, In Trans. on PAMI.

论文(paper)

在单张静态彩色图像上进行动作识别的论文不是很多, 可以从两方面进行区分:

1 传统模型

2 深度模型.

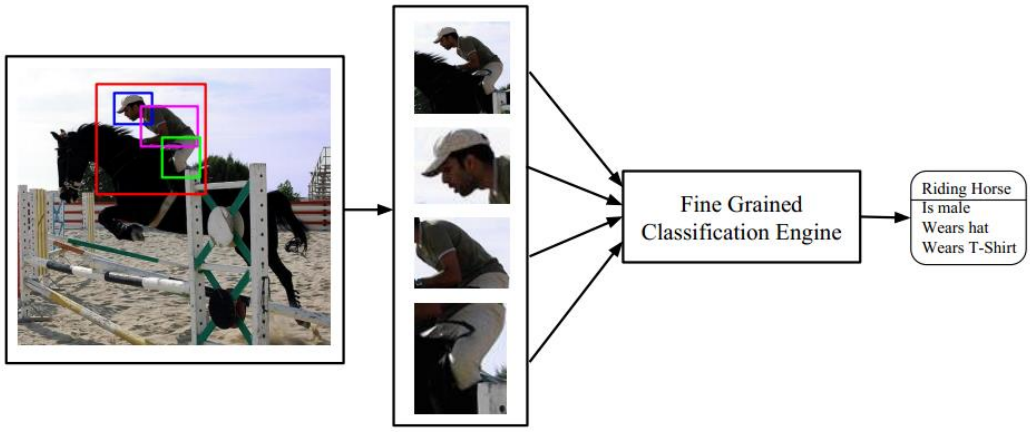
下面分布从这两方面对一些经典的近年来的论文进行概要性的介绍.

深度模型

1 Actions and Attributes from Wholes and Parts. (2014)

Method:

该论文利用[4]对给定的 human bounding box(这个可以利用 human detector 获得, 如可以训练一个[2]的 R-CNN human detector)进行 parts 的检测, 然后将 human bounding box 和 parts bounding boxes 的图像块作为 CNN(类似于[2]的 R-CNN)的输入来提出 fc7 的特征, 最后这些特征作为 SVM 的输入, 来进行 action classification 和 attribute classification 的任务, 其框架如图 1(a)和图 1(b)所示:

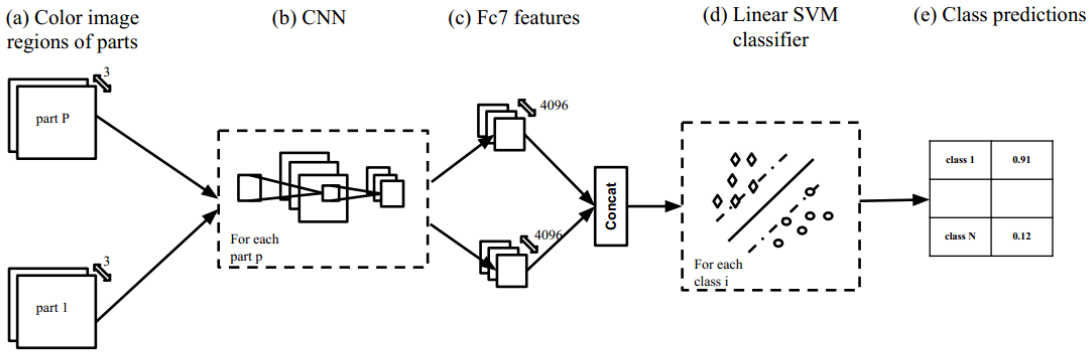


(a) Given an instance hypothesis, we detect parts

(b) The instance and its parts are fed into our classification engine

Schematic overview of our overall approach. (a) Given an R-CNN person detection (red box), we detect parts using a novel, deep version of poselets (Section 3). (b) The detected whole-person and part bouding boxes are input into a fine-grained classification engine to produce predictions for actions and attributes (Section 4).

图 1(a) 整体框架



Schematic overview of our approach for fine grained classification using parts. (a) We consider regions of part activations. (b) Each part is forward propagated through a CNN. (c) The output is the fc7 feature vector for each input. (d) The features are concatenated and fed into linear SVM classifiers. (e) The classifiers produce scores for each class.

图 1(b) 测试阶段的流程

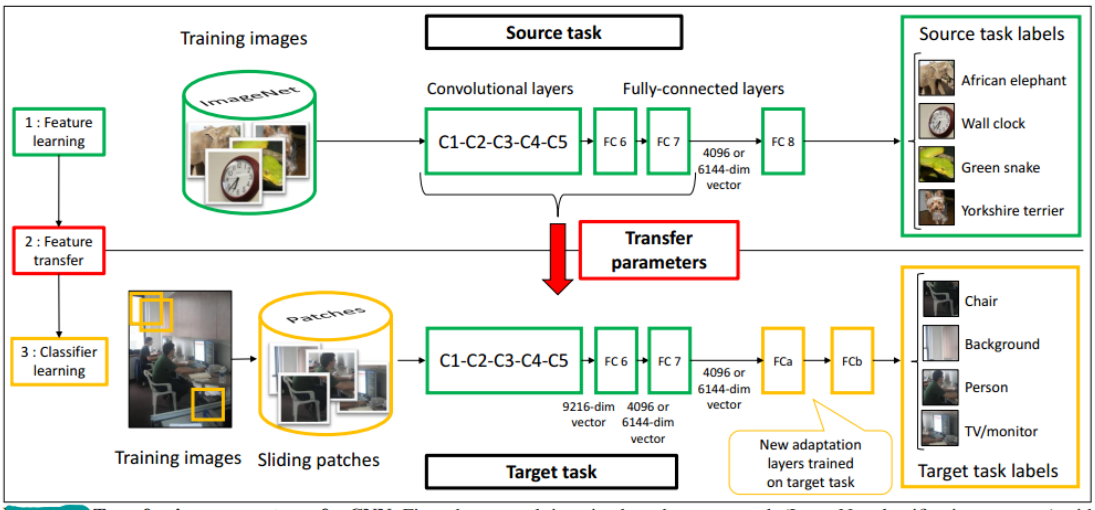
Dataset:

PASCAL VOC 2012 action dataset

2 Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks (2014 CVPR)

Method:

该论文通过将在 ImageNet 数据集上训练好的 AlexNet[3]的模型, 来初始化论文里定义好的 CNN 模型, 并进行 fine-tuning. 如图 2 所示, 在 Pascal Voc 的数据集上进行 Object localization 和 Action recognition 的任务.



Transferring parameters of a CNN. First, the network is trained on the source task (ImageNet classification, top row) with a large amount of available labelled images. Pre-trained parameters of the internal layers of the network (C1-FC7) are then transferred to the target tasks (Pascal VOC object or action classification, bottom row). To compensate for the different image statistics (type of objects, typical viewpoints, imaging conditions) of the source and target data we add an adaptation layer (fully connected layers FCa and FCb) and train them on the labelled data of the target task.

图 2 transferring parameters of a CNN

Dataset:

Pascal VOC 2007 datasets, Pascal VOC 2012 datasets

3 R-CNNs for Pose Estimation and Action Detection (2014 CVPR)

Method:

该论文利用 MCG[1]来产生 region proposals, 基于 AlexNet[3]的架构来定义 R-CNN 模型[2], 并用 AlextNet[3]的模型来 fine-tuning. 该 R-CNN 模型同时进行多任务的学习(包括 human detection, pose estimation and action classification). 具体模型如图 3 所示:

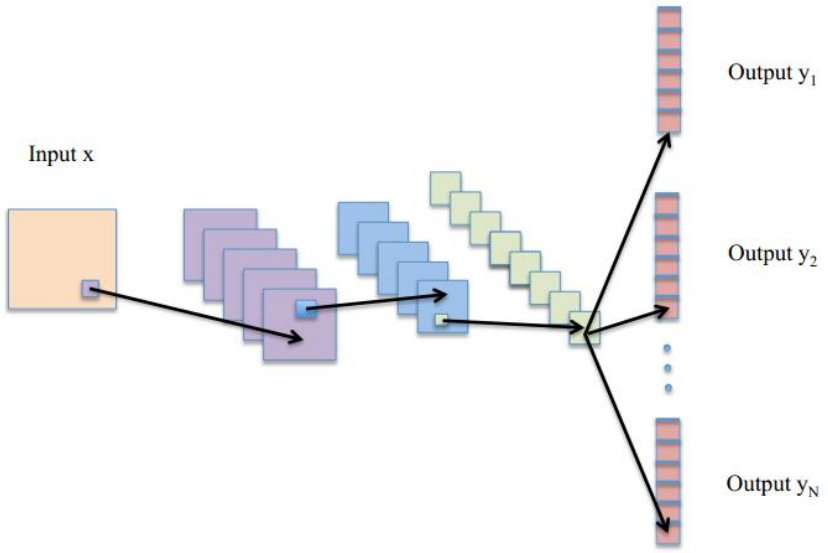


图 3 RCNN 框架

Dataset:

Pascal VOC 2012 action dataset

传统模型 (注: 这里的传统模型相对于深度模型而言)

1 Recognizing Actions from Still Images (CVPR 2008)

Method:

该论文利用[6]里面的 model 来对图像进行 human pose 的 estimation, 然后在 human pose estimation 的结果(是一个 sparse probability map)上, 根据[7]里面的方法提取 HOGRs(Histogram of Oriented Rectangles)的特征, 接着利用 LDA 对特征进行降维, 最后采用 one-vs-all 的方式训练 SVM 来对动作进行分类.

Dataset:

没有公开论文里面的数据集, 论文里面的数据集是由作者收集到的, 一共有 467 张静态彩色图, 由 6 个不同的动作类别组成:

Running
Walking
Catching
Throwing
Crouching
Kicking

2 Learning person-object interactions for action recognition in still images (NIPS 2011)

Method:

该论文主要是对 person-object interaction 的关系进行建模,来进行动作的识别, 首先利用[8]和[9]对图像进行 object 和 person 的 localization, 得到一个关于(x, y, s)的 D 个 cell 的金字塔(每个检测子都有这样一个金字塔), 然后根据(x, y, s)来对 pyramid 上的每个 cell 的 person-object interact 进行 pair-wise interaction 建模, 这样可以得到每个 pair-wise interaction 的 score, 并通过 SVM 来选择合适的 M 个具有判别性的 pair-wise interaction. 最后对每个类别, 每个 cell 的每个 pair-wise 的 score 组成一个长向量, 作为最后的动作识别分类器的特征, 来对动作进行分类.

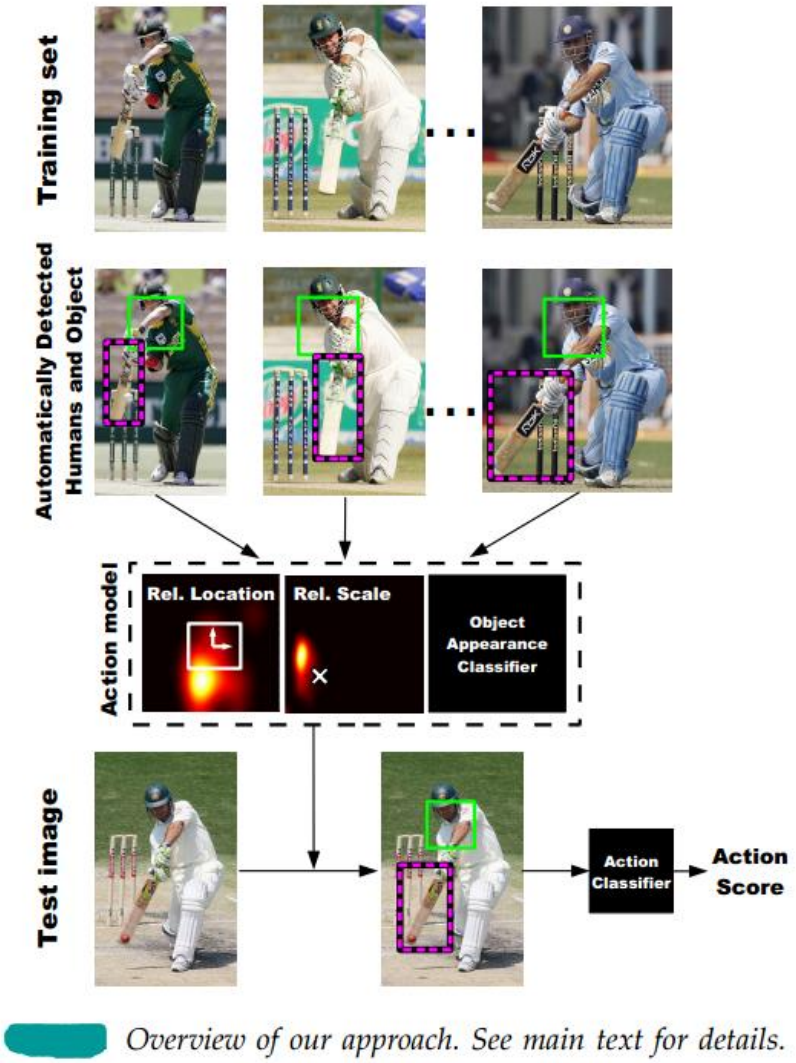
Dataset:

Willow-action dataset, PASCAL VOC 2010 action classification dataset.

3 Weakly supervised learning of interactions between humans and objects (PAMI 2012)

Method:

该论文主要是对 person-object interaction 进行建模, 首先利用 human detector 对图像里的 subject 进行检查, 然后围绕这检测到 subject 和利用 object detector 来检测到 interacted 的 object, 然后基于检测到的 subject 和 object 进行特征提取(这个没怎么看), 将特征作为训练好的 action classifiers(即多分类的 SVM)的输入来进行动作的识别. 整体的框架如图 3 所示.



Overview of our approach. See main text for details.

图 3(a)整体框架

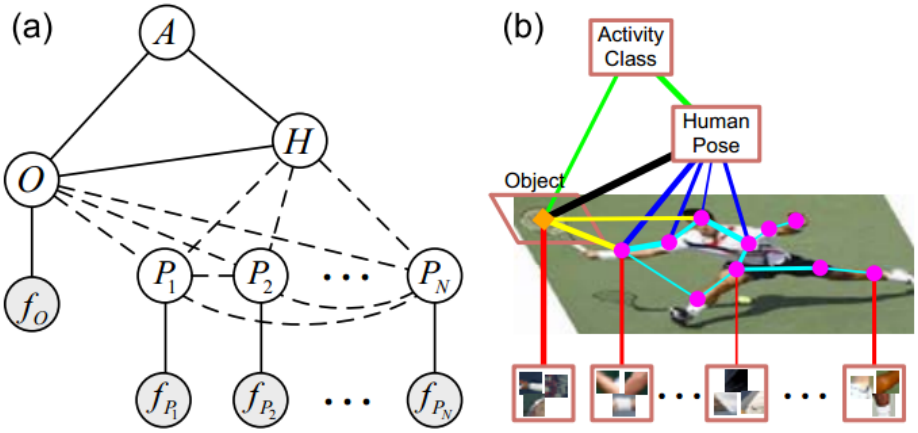
Dataset:

sports action dataset(six-class sports dataset), the PASCAL Action 2010 dataset], TBH dataset (TBH 数据集没有公开)

4 Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities (CVPR 2010)

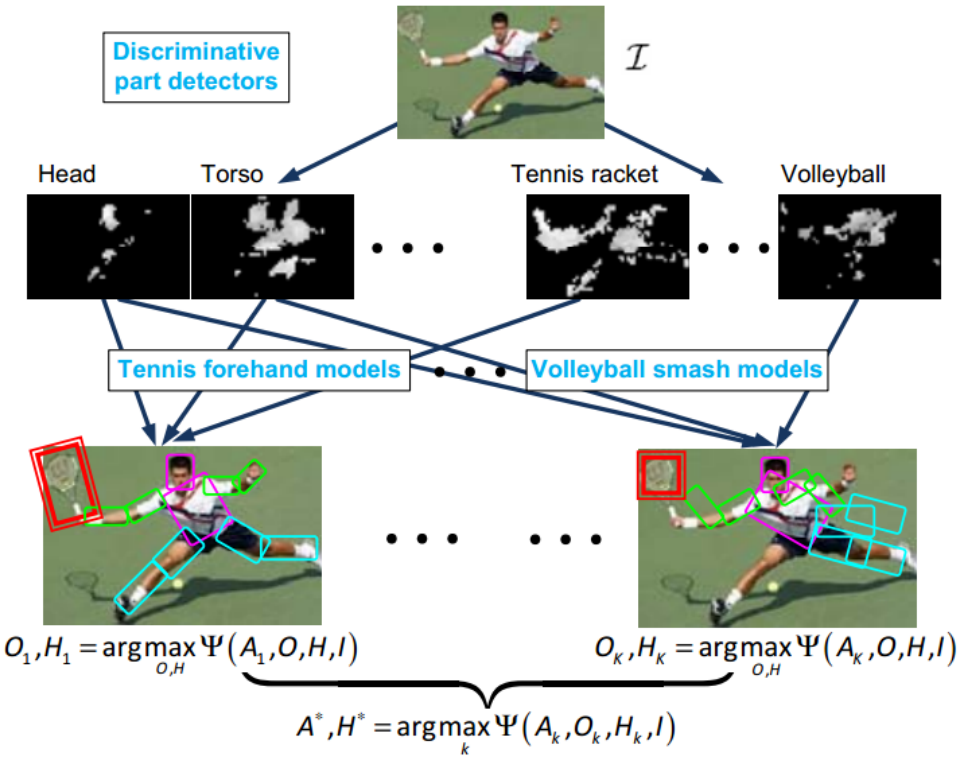
Method:

该论文主要是对 human-object 的 mutual context 进行建模, 利用 object detector 和 human detector 对图像中的 subject 和 object 进行检查(找出 object 和 subject 的 parts), 然后根据 parts 之间的连接(connectivity), 利用 random field model 来对(类别, 物体, 人)之间的关系进行建模, 在测试的时候, 通过搜索策略来最大化 randomfield model 的目标函数, 从而找出最可能的动作类别, 其模型和测试阶段如图 4(a)和图 4(b)所示.



(a) A graphical illustration of our model. The edges represented by dashed lines indicate that their connectivity will be obtained by structure learning. A denotes an HOI activity class, H the human pose class, P a body part, and O the object. f_O and f_P's are image appearance information of O and P respectively. **(b)** Illustration of our model on an image of a human playing tennis. Different types of potentials are denoted by lines with different colors. Line widths represent the importance of the potentials for the human-object interaction of playing tennis.

图 4(a)模型框架



The framework of our inference method. Given an input image $\mathcal{I}$, the inference results are: (1) object detection results O_k (e.g. O_1 is the tennis racket detection result); (2) human pose estimation result H^*; (3) activity classification result A^*.

图 4(b) 测试过程

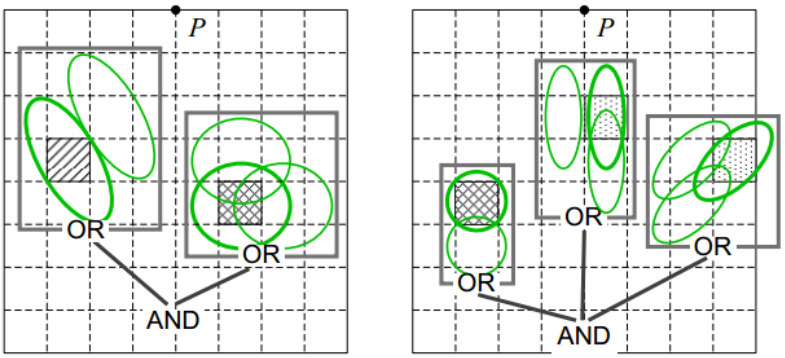
Dataset:

sports action dataset(six-class sports dataset)

5 Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions (CVPR 2010)

Method:

该论文通过定义了一种多个图像块的 grouplet 来表达图像特征，也就是 grouplet(如图 5 所示), 可以捕获到一个图像的结构化的信息, 如图像块的外观信息和图像块之间的 spatial configuration. 在论文中, 通过找出图像里面的多个 grouplet, 即 grouplets, 然后从 grouplets 中提出特征, 将这些特征作为产生式模型或者判别式模型(如 SVM)的输入, 来产生图像的动作类别.



Two examples of grouplets: **left** is a size 2 grouplet; and **right** is a size 3 grouplet. Each grouplet lives in an image space where P indicates a reference location. Each grouplet is composed of a set of feature units. A feature unit, whose visual appearance is denoted by a shaded square patch, can shift around in a local neighborhood (indicated by smaller rectangular boxes). An ellipse surrounding the center of a feature unit indicates the spatial extent of the feature. Within the neighborhood, an OR operation is applied to select the feature unit that has the strongest signal (v, see Sec.4.1), indicated by the ellipse of thicker lines. An AND operation collects all feature units to form the grouplet.

图 5 grouplet

Dataset:

PPMI (people-playing-musical-instrument)

6 On Recognizing Actions in Still Images via Multiple Features (ECCV 2012)

Method:

该论文利用[10]的方法来产生 multiple object 的 hypotheses, 然后利用 MIL(Multiple Instance Learning)来从 multiple object 的 hypotheses 中产生 candidate related object hypotheses. 最后从这些 candidate object hypotheses 中提取多种特征(如 facial feature, HOG, BOW 等), 最后为每一种特征训练一个 one-vs-rest 的 SVM. 通过将每种特征的 SVM 的 score 进行线性相加得到每个类别最后的 score 来进行动作识别.

Dataset:

Stanford 40 Actions dataset

7 Action Recognition from a Distributed Resprentation of Pose and Appearance (CVPR 2010)

Method:

该论文利用[9, 10]的方法来训练与动作类别相关/明确的 poselets 模型来检测图像中的 subject, 以及用 poselet 的方式来检测与 subject 最接近(相关)的 object, 然后在这些 poselets 上提取外观特征和空间几何约束特征, 最后将这些特征做为 SVM 的输入, 来训练动作分类的 SVM 模型.

Dataset:

Pascal Voc 2010 action dataset

8 Recognizing Human Actions from Still Images with Latent Poses (CVPR 2010)

Method:

不同于之前的将 pose estimation 和 action recognition 分开的论文, 该论文提出一个新的方法来进行动作识别, 该方法将图像中的人的 pose 作为隐变量, 来直接利用 pose 的信息来帮助 action recognition.

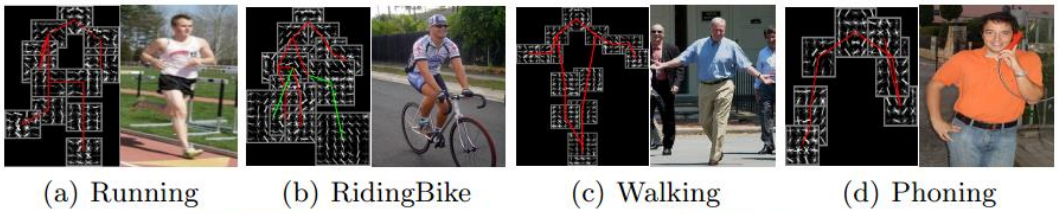
Dataset:

five action categories from [13]

9 Detecting Actions, Poses, and Objects with Relational Phraselets (ECCV2012)

Method:

该论文提出一个新的方法来组合 Skeleton model, Poselets 和 Visual Phrse 三种模型来捕获 object 和 person 的 articulation, 以及 object 和 person 之间的 interaction, 来进行 action recognition 和 pose estimation. 其模型的可视化如图 9 所示.



(a) Running (b) RidingBike (c) Walking (d) Phoning

(e) RidingHorse (f) UsingComp. (g) TakingPhoto (h) Jumping

Visualizations of our learned models and tree-structured relations. Our activity-specific tree connects part templates spanning both, the human and the object. Red edges connect parts of the human to each other. Green edges connect parts of an object to each other and to the human. Note that we are showing one (out of an exponential number of) combinations of local templates for each activity. For example, the selected phraselet mixtures in **(e)** correspond to a left-facing horse, but the same model generates other views by swapping out different mixtures at different spatial locations (as shown in Fig. 1).

图 9 模型可视化

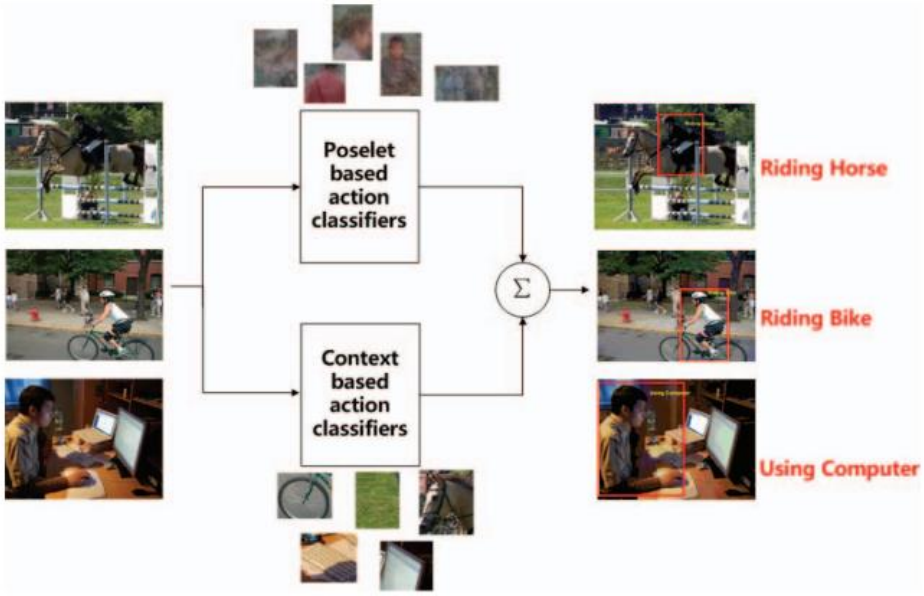
Dataset:

PASCAL 2011 action classification

10 Action Recognition in Still Images Using a Combination of Human Pose and Context Information. (2012 ICIP)

Method:

该论文提出了一种新的动作识别的方法, 该方法利用 Poselet-based action classifiers 来获取 pose information 的 Poselet Activatoin Vector 作为特征, 和利用 spar coding 的方式来获取 contextual information 作为特征训练 context-based action classifiers, 最后利用 pose information 和 contextual information 来进行动作识别. 如图 10 所示.



We use pose and context information together to recognize human actions in still images. We first train poselet-based action classifiers and context-based action classifiers for each action. Then given a test image, the probability outputs from the two classifiers are summed up as the confidence of the image belonging to each action. The action with the highest confidence is selected as the predicted action label.

图 10 系统可视化

Dataset:

PASCAL VOC 2010 Dataset, PASCAL VOC 2011 Dataset, Willow Action Dataset

11 Classifying Actions and Measuring Action Similarity by Modeling the Mutual Context of Objects and Human Poses (ICML 2011)

Method:

该论文通过使用共同上下文模型(mutual context model)来同时对 pose, object, 和 person-object 的 interaction 进行建模, 来对动作进行识别.

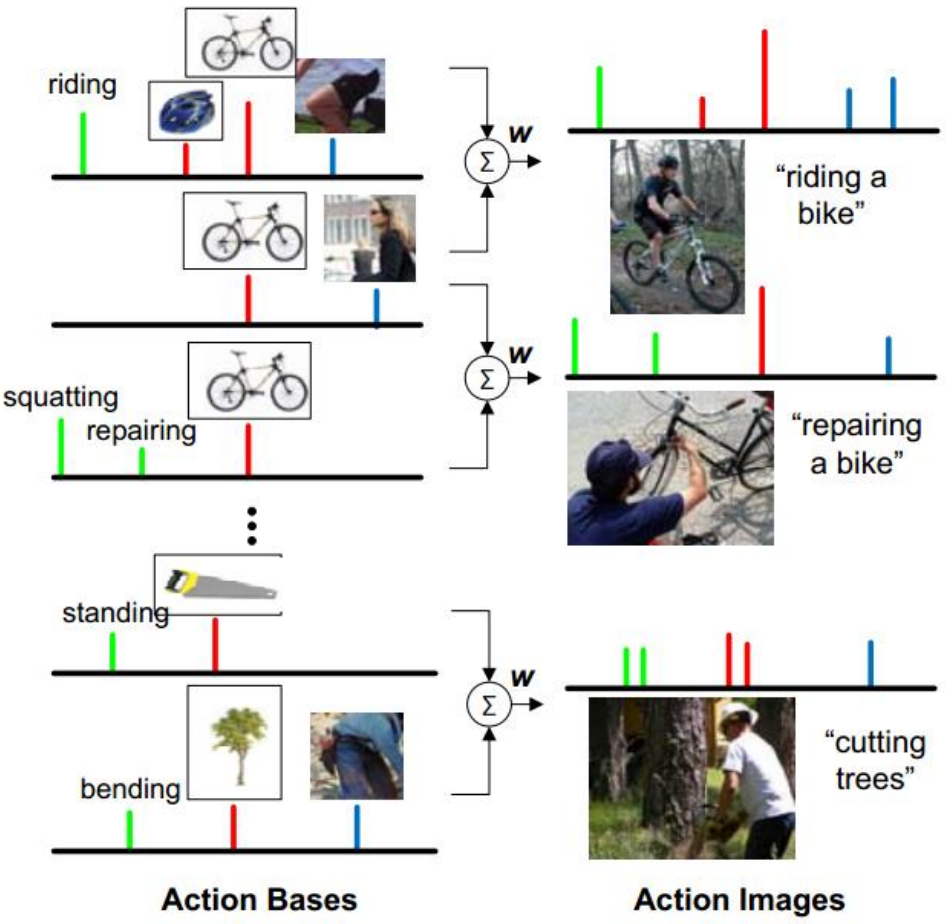
Dataset:

Six-Class Sports Dataset

12 Human Action Recognition by Learning Bases of Action Attributes and Parts (ICCV 2011)

Method:

该论文提出了一种利用 action 的 attributes 和 parts 的方式, 来训练与动作类别相关的 sparse bases. 然后利用训练好的 sparse bases 来重构图像中的 attributes 和 parts, 最后根据重构后的 attributes 和 parts, 来决定动作类别, 如图 12 所示



We use attributes (verb related properties) and parts (objects and poselets [2]) to model action images. Given a large number of image attributes and parts, we learn a number of sparse action bases, where each basis encodes the interactions between some highly related attributes, objects, and poselets. The attributes and parts of an image can be reconstructed from a sparse weighted summation of those bases. The colored bars indicate different attributes and parts, where the color code is: green - attribute, red - object, blue - poselet. The height of a bar reflects the importance of this attribute or part in the corresponding basis.

图 12

Dataset:

PASCAL action dataset, Stanford 40 Actions dataset

13 Recognizing human actions in still images: a study of bag-of-features and part-based representations (BMVC 2010)

Method:

该论文利用 bag-of-features 的特征，和 part-based latent SVM[15]的方式来提取 HOG 特征，分别利用这两种特征来训练基于 SVM 的 action classifiers，并利用这些 action

classifiers 来单独或者组合的方式, 进行 action classification.

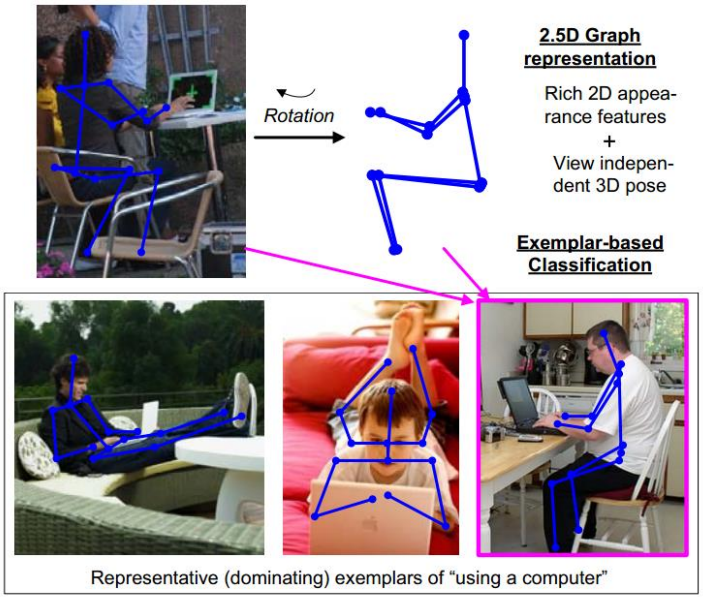
Dataset:

Willow actions database, Six-Class Sports dataset, PPMI (people-playing-musical-instrument)

14 ActionRecognitionwithExemplarBased 2.5DGraphMatching (ECCV 2012)

Method:

该论文提出了一种新颖的 2.5D 的动作识别的 representation, 即 2.5D graph. 2.5D graph 由 human body 的 key-points 组成的节点和这些节点之间表示空间关系的边组成. 每个节点由一个 view-independent 的 3D 坐标和局部的 2D 外观特征组成. 论文利用 exemplars based action classification 的方法来根据图像的 2.5D graph 进行动作识别. 如图 14 所示.



An overview of our action recognition algorithm. We represent an action image as a 2.5D graph consisting of view-independent 3D pose and 2D appearance features. In recognition, the 2.5D graph is matched with a set of exemplar graphs for each action class, allowing more robust handling of within-action variations.

图 14

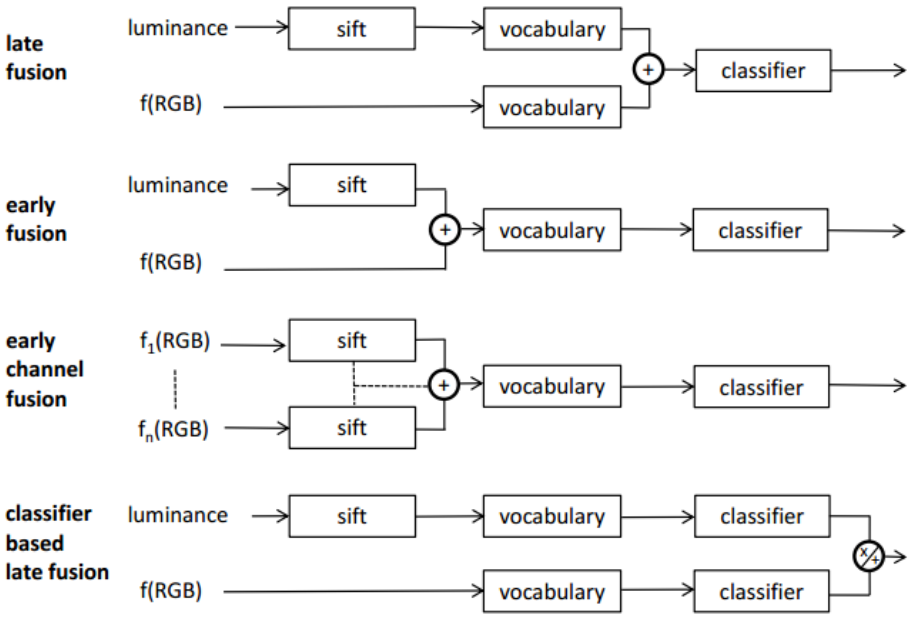
Dataset:

PPMI dataset, Pascal Voc 2011 action dataset

15 Coloring Action Recognition in Still Images (IJCV 2013)

Method:

该论文利用 color descriptors 和 shape descriptors, 来进行 BOW 的特征提取, 同时探究了不同特征之间的 fusion, 来训练动作识别分类器. 其 pipeline 如图 15 所示.



Pipelines for four different fusion methods. The fusion between color and shape is indicated by a 'plus' in case of concatenation of vectors or vocabulary histograms. In the case of classifier based fusion, the encircled multiplication and sum symbols refer to the two methods of classifier fusion investigated: summation and multiplication, respectively, of their outputs. The function $f(RGB)$ refers to a mapping of RGB values to another color-space representation. The "vocabulary" modules refer to vocabulary assignment and have histograms as output. Methods which perform fusion before vocabulary assignment are called early fusion methods, otherwise they are late fusion approaches.

图 15

Dataset:

Willow, PASCAL VOC 2010, Stanford-40

16 Semantic Pyramids for Gender and Action Recognition (TIP 2014)

Method:

该论文提出了一种语义金字塔的方法来弄 pose normalization. 该方法能够自动地从 face, upper-body 和 full-body 的 region 上提出基于 BOW 的 spatial pyramid representation, 组成一个 single long feature vector, 来进行动作识别.

Dataset:

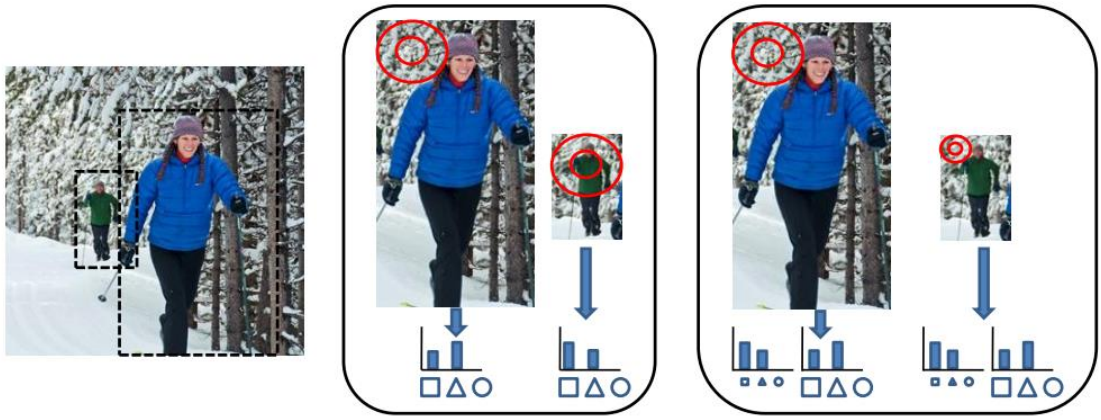
Willow, Stanford 40, Pascal Voc 2010 actions dataset, Six-Class Sports dataset

17 Scale Coding Bag-of-Words for Action Recognition (ICPR 2014)

Method:

该论文提出了一种尺度变化的特征编码方式, 区别以往的尺度不变的 BOW 方式来获取给定图像或者图像中 bounding box 的特征, 将不同尺度上得到的特征

(histograms)连接起来(如图 17 所示)，作为活动分类器的特征.



Scale coding: (left) input image, superimposed bounding boxes indicate persons performing an action; (middle) in standard scale coding the scale is independent of the object size (red circles show the extracted feature scales), and they are all assembled in a single histogram per image; (right) our proposal of relative scale coding adapts to the bounding box of the object. This ensures that similar structures (such as hands and ski poles) are captured at the same scale independent of the bounding box size. The features are represented in several concatenated histograms which collect a range of feature scales.

图 17 多尺度的特征编码方式

Dataset:

Willow action dataset,

18 Integrating Randomization and Discrimination for Classifying Human-Object Interaction Activities (Human-Centered Social Media Analytics 2014)

Method:

该论文提出了一种发现可用作 fine-grained categorization tasks 的具有高度判别性的 image patches 和 pairs of patches 的判别性的决策树的随机森林. 随机森林通过使用在每个节点上的强分类器和组合决策树上不同深度下的信息来，进行活动分类. 如图 18(a)和图 18(b)所示

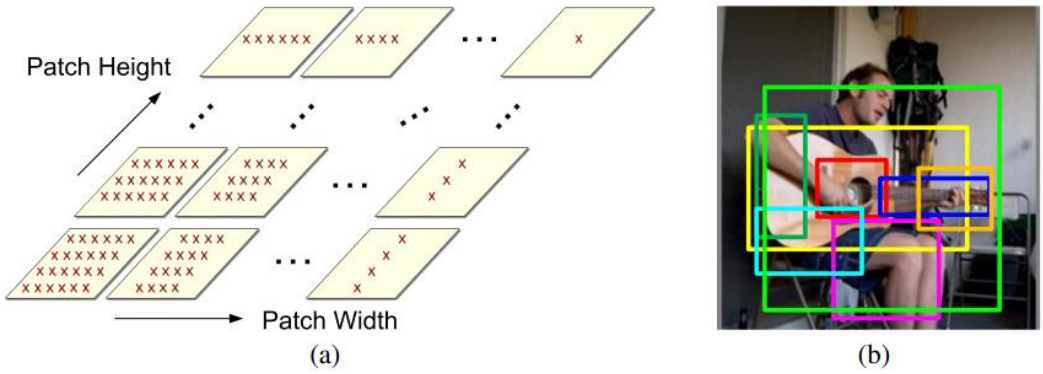
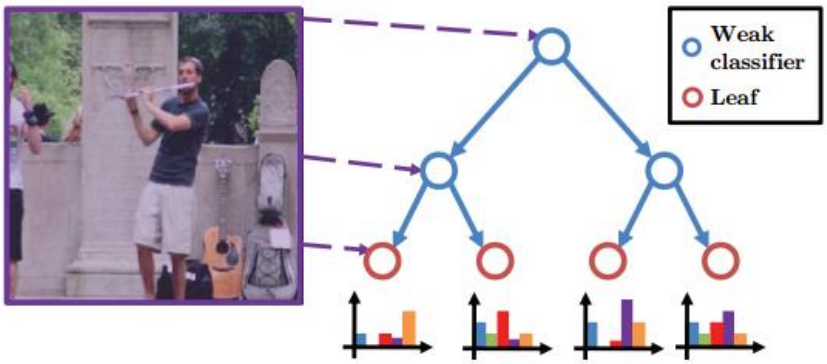
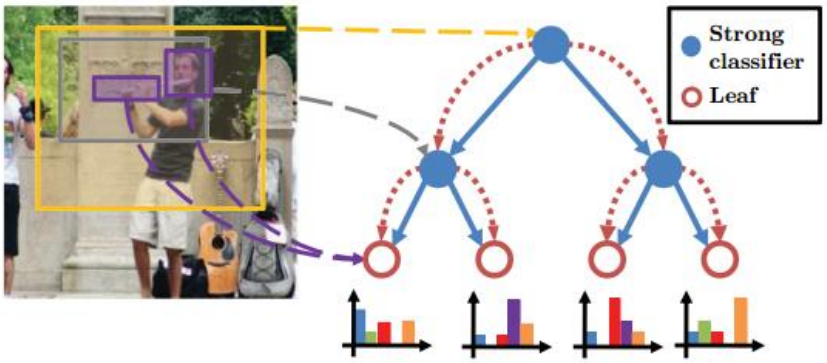


Illustration of the proposed dense sampling space. (a) We densely sample rectangular image patches with varying widths and heights. The regions are closely located and have significant overlaps. The red × denote the centers of the patches, and the arrows indicate the increment of the patch width or height. (b) Illustration of some image patches that may be discriminative for "playing-guitar". All those patches can be sampled from our dense sampling space.

图 18(a) sampled patches



(a) Conventional random decision tree.



(b) The proposed discriminative decision tree.

Comparison of conventional random decision trees with our discriminative decision trees. Solid blue arrows show binary splits of the data. Dotted lines from the shaded image regions indicate the region used at each node. Conventional decision trees use information from the entire image at each node, which encodes no spatial or structural information, while our decision trees sample single or multiple image regions from the dense sampling space (Figure 2(a)). The histograms below the leaf nodes illustrate the posterior probability distribution $P_{t,l}(c)$ (Section 4.1). In (b), dotted red arrows between nodes show our nested tree structure that allows information to flow in a top-down manner. Our approach uses strong classifiers in each node (Section 4.3), while the conventional method uses weak classifiers.

图 18(b) random decision trees

Dataset:

PPMI

Reference

1 P. Arbel´aez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In CVPR, 2014.

2 R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. InCVPR, 2014.

3 A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

4 R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. arXiv preprint arXiv:1409.5403, 2014.

5 N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In ECCV, 2014.
6 D. Ramanan. Learning to parse images of articulated bodies. In NIPS, 2006.
7 N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In Human Motion Workshop LNCS 4814 , pages 271–284, 2007.
8 P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE PAMI, 2009.
9 S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. InCVPR, 2011.
10 Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, USA (2010)
11 L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In ECCV, 2010.
12 L. Bourdev, S. and J. Malik. Poselet: Body part detectors trained using 3d human pose annotations. In ICCV, 2009.
13 N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning actions from the web. InICCV, 2009.
14 J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV, 73(2): 213–238, 2007.
15 P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models.IEEE PAMI, 2009.