An Analysis of Action Recognition Datasets for Language and Vision Tasks

将视觉任务语言任务结合

----行为识别（action recognition）

将视觉任务语言任务结合

----行为识别(action recognition)

应用场景：

- 图像标注（image annotation）
- 情境理解（scene understanding）
- 图像检索（video/image retrieve）
- 人机交互（human-computer interaction）

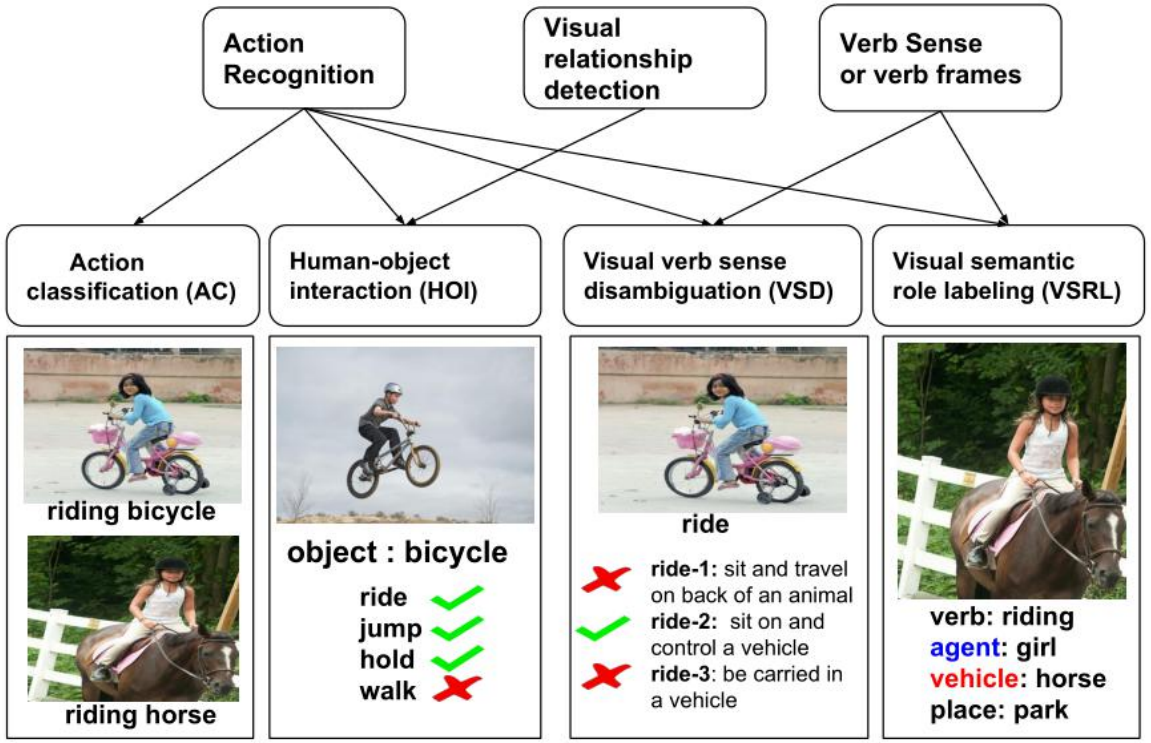
行为识别大多基于视频一类的动态图像，通过提取时空特征进行分析，当考虑到如下图所示的静态图像中无法提供时空上的特征表示，从图片进行行为识别则显得更具挑战性

riding horse　　running　　playing guitar　　jumping

Figure 1: Examples of actions in still images

实现目标：对图中的动作进行细粒度的语义、语法分析

- 行为划分（AC）
- 确定交互的人-客体（HOI）
- 视觉性动词词义消歧（VSD）
- 视觉语义角色标注（VSRL）



行为划分(Action Classification)

在小规模的数据集上，使用预先确定的动词短语或动宾词组作为行为标签进行训练。基于一张给定的图片仅与单一行为相关的假想，模型会将各个图片冠以互斥的行为标签

在建立模型对图片进行理解上有所帮助，但是使用的许多方法无法泛化推广到更大规模的数据集上，同时互斥标签的理念与实际不符，有些动作常常是同时出现于某一情境中

holding

playing

guitar

确定交互关系(Determining Human-Object Interaction)

为解决行为划分的互斥标签问题而生，目的是确定图像中以人为主体、与相应客体的所有交互关系对

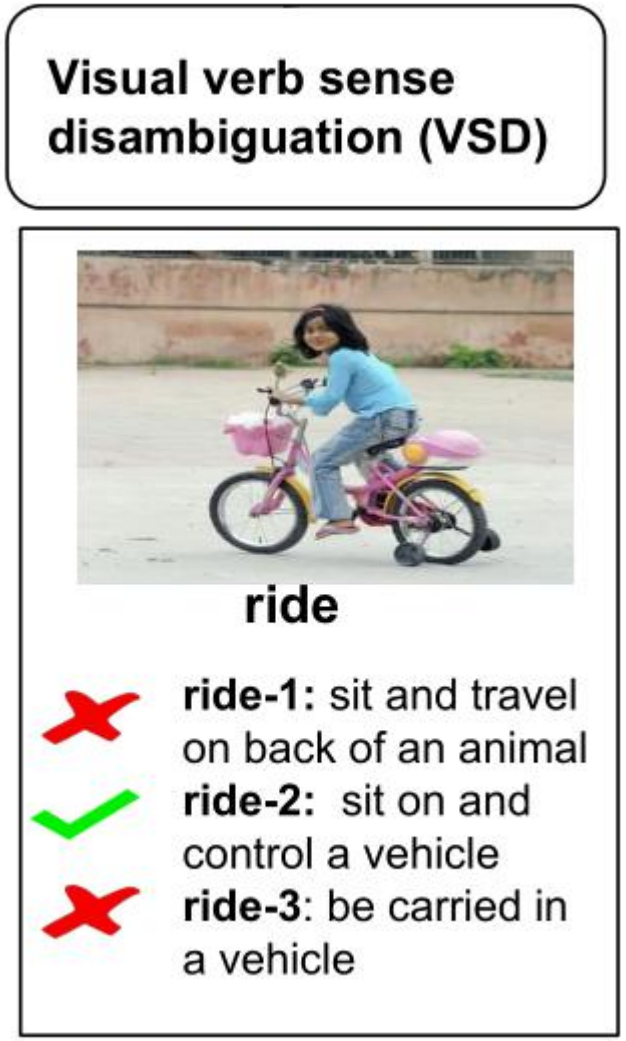
AC与HOI存在的问题

- 多义动词的使用会引起歧义
- 某些表达方式存在共有的概括表达
- 不同表达方式可以表达同种意义

由此提出在动词词义层面进行行为分析

视觉性动词词义消歧(Visual Verb Sense Disambiguation)

使用现有的词库资源进行词义-图片对标记，能够区分动词不同词义的应用，但是不能够定位行为对应的客体



视觉语义角色标注(Visual Semantic Role Labeling)

以动词为基础搭建框架，
将客体填入语义角色的位置

即确定主-谓-宾
（subject-verb-object）结构

Dataset	Task	#L	#V	Obj	Imgs	Sen	Des	Cln	ML	Resource	Example Labels
Ikizler (Ikizler et al., 2008)	AC	6	6	0	467	N	N	Y	N	–	running, walking
Sports Dataset (Gupta et al., 2009)	AC	6	6	4	300	N	N	Y	N	–	tennis serve, cricket bowling
Willow (Delaitre et al., 2010)	AC	7	6	5	986	N	N	Y	Y	–	riding bike, photographing
PPMI (Yao and Fei-Fei, 2010)	AC	24	2	12	4.8k	N	N	Y	N	–	play guitar, hold violin
Stanford 40 Actions (Yao et al., 2011)	AC	40	33	31	9.5k	N	N	Y	N	–	cut vegetables, ride horse
PASCAL 2012 (Everingham et al., 2015)	AC	11	9	6	4.5k	N	N	Y	Y	–	riding bike, riding horse
89 Actions (Le et al., 2013)	AC	89	36	19	2k	N	N	Y	N	–	ride bike, fix bike
MPII Human Pose (Andriluka et al., 2014)	AC	410	–	66	40.5k	N	N	Y	N	–	riding car, hair styling
TUHOI (Le et al., 2014)	HOI	2974	–	189	10.8k	N	N	Y	Y	–	sit on chair, play with dog
COCO-a (Ronchi and Perona, 2015)	HOI	–	140	80	10k	N	Y	Y	Y	VerbNet	walk bike, hold bike
Google Images (Ramanathan et al., 2015)	AC	2880	–	–	102k	N	N	N	N	–	riding horse, riding camel
HICO (Chao et al., 2015)	HOI	600	111	80	47k	Y	N	Y	Y	WordNet	ride#v#1 bike; hold#v#2 bike
VCOCO-SRL (Gupta and Malik, 2015)	VSRL	–	26	48	10k	N	Y	Y	Y	–	verb: hit; instr: bat; obj: ball
imSitu (Yatskar et al., 2016)	VSRL	–	504	11k	126k	Y	N	Y	N	FrameNet	verb: ride; agent: girl#n#2
										WordNet	vehicle: bike#n#1; place: road#n#2
VerSe (Gella et al., 2016)	VSD	163	90	–	3.5k	Y	Y	Y	N	OntoNotes	ride.v.01, play.v.02
Visual Genome (Krishna et al., 2016)	VRD	42.3k	–	33.8k	108k	N	N	Y	Y	–	man playing frisbee

Table 1: Comparison of various existing action recognition datasets. #L denotes number of action labels in the dataset; #V denotes number of verbs covered in the dataset; Obj indicates number of objects annotated; Sen indicates whether sense ambiguity is explicitly handled; Des indicates whether image descriptions are included; Cln indicates whether dataset is manually verified; ML indicates the possibility of multiple labels per image; Resource indicates linguistic resource used to label actions.

早期数据集的变化(AC)

- 借助图像视角、背景、分辨率来分析

  eg：throwing、running

- 捕捉动作、姿态、身体部位的区别以建立特征模型

  eg： tennis serve 、cricket bowling

- 分辨直观的客体信息

  eg：riding horse、riding bike

下一阶段

- 覆盖面更加广泛
- 使用已有的语言学资源

  eg：VerbNet、OntoNote、FrameNet
- 涉及不能在图像中直观显示的动词或动词词义

  eg： playing with emotions、playing instruments

这些数据集的发展呈现两种态势：

- 人工筛选标签或集中在特定的领域
- 以特定客体类型集中选取行为（对图像描述和QA任务数据集的分析，表明名词的分布要高于其它词类）

相关行为识别模型

AC或HOI这样的任务一般需要较高层次的图像信息，例如人型或人体部位、客体、情境

这些高层次的信息要从低层次的特征中提炼，如尺度不变特征变换（Scale-invariant feature transform，SIFT）、方向梯度直方图(Histogram of Oriented Gradient, HOG)、Gist特征，同时也会用到主客的相对位置、角度等信息

最近的一些模型开始使用基于端到端的CNN结构

相关行为识别模型

大部分模型仅基于视觉信息，同时也出现了与其它方向结合的模型

- 借助语言学知识
- 使用在大量文本语料中训练出的词向量

部分研究表明，与图像相关联的文本资源中训练出的embedding的使用，有助于动词语义消歧，并对视觉信息的隐藏信息有所补充

思考：将语言学资源与视觉信息结合

- 可以作为文本语义消歧的延伸
- 视觉资源有助于一些NLP问题（如bilingual lexicon induction）
- 两者结合有助于多语言问题的解决，如源语言到目标语言的异义异词和异义同词

当前主要问题

- 覆盖领域相对狭隘
- 使用的资源互不一致