(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106845351 A
(43)申请公布日 2017.06.13

(21)申请号 201611193290.1

(22)申请日 2016.12.21

(66)本国优先权数据
201610316012.4 2016.05.13 CN

(71)申请人 苏州大学
地址 215123 江苏省苏州市苏州工业园区仁爱路199号

(72)发明人 刘纯平 葛瑞 季怡 刘海宾 龚声蓉

(74)专利代理机构 苏州创元专利商标事务所有限公司 32103
代理人 陶海锋

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)

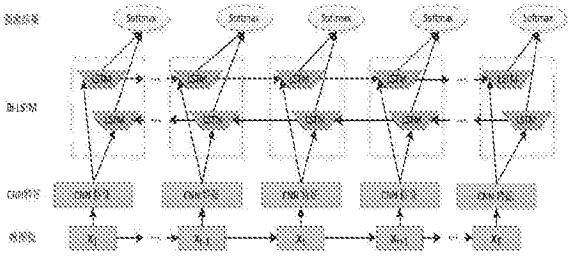
权利要求书1页 说明书6页 附图2页

(54)发明名称

一种用于视频的基于双向长短时记忆单元的行为识别方法

(57)摘要

本发明公开了一种用于视频的基于双向长短时记忆单元的行为识别方法，包括：(1)输入视频序列，提取视频序列中的RBG帧序列和光流图像；(2)分别训练RGB图像深度卷积网络和光流图像深度卷积网络；(3)提取网络的多层特征，其中至少提取第三卷积层、第五卷积层、第七全连接层的特征；对卷积层特征进行和池化；(4)对采用双向长短时记忆单元构建的递归神经网络进行训练，得到视频每帧的概率矩阵；(5)对每个概率矩阵取平均，最后融合光流帧和RGB帧的概率矩阵，取概率最大的类作为最后的分类结果，由此实现行为识别。本发明采用多层深度学习的特征代替传统的人工特征，不同层的深度特征表征了不同的信息，多层特征的组合可以提高分类的准确率；通过采用双向长短时记忆捕获时间信息，获得更多的时域结构信息，提高了行为识别的效果。



CN 106845351 A

1.一种用于视频的基于双向长短时记忆单元的行为识别方法,包括以下步骤:

(1)输入视频序列,提取所述视频序列中的RBG帧序列和光流图像;

(2)训练深度卷积网络:分别训练RGB图像深度卷积网络和光流图像深度卷积网络;

(3) 提取网络的多层特征,其中至少提取第三卷积层、第五卷积层、第七全连接层的特征;对全连接层得到一个固定大小的向量;对卷积层特征进行和池化,加入时间信息;

(4) 利用从卷积神经网络得到的各层特征向量,对采用双向长短时记忆单元构建的递归神经网络进行训练,并输入测试集利用Softmax得到视频中光流帧每帧的概率矩阵和RGB帧每帧的概率矩阵;

(5) 分别对视频中所有光流帧的概率矩阵取平均,对视频中所有RGB帧的概率矩阵取平均,最后融合光流帧和RGB帧的概率矩阵,取概率最大的类作为最后的分类结果,由此实现行为识别。

2.根据权利要求1所述的用于视频的基于双向长短时记忆单元的行为识别方法,其特征在于:步骤(2)中,选择第三卷积层、第五卷积层和第七全连接层作为特征表达。

3.根据权利要求1所述的用于视频的基于双向长短时记忆单元的行为识别方法,其特征在于,所述和池化为:

在时间t第m层特征图是F_m^t,使用下面的公式进行池化:

$$Desc = \sum_{j=1}^{N} F_m^{t_j}(x,y)$$

其中,j是池化帧,N是时域池化范围,最后描述子的维度是$H_s \times W_s \times C \times T$,$H_s$是特征图的高度,$W_s$是特征图的宽度,C是特征图通道个数,T是视频的总帧数。

4.根据权利要求3所述的用于视频的基于双向长短时记忆单元的行为识别方法,其特征在于:在所述和池化后步骤(4)之前,使用空间金字塔最大池化,使用两层金字塔,得到维度为定长的特征向量。

5.根据权利要求3所述的用于视频的基于双向长短时记忆单元的行为识别方法,其特征在于:N为15。

一种用于视频的基于双向长短时记忆单元的行为识别方法

技术领域

[0001] 本发明涉及一种视频处理方法，具体涉及一种自动识别视频中人物的行为的方法。

背景技术

[0002] 行为识别是指通过提取视频或图像序列中的特征信息对目标的行为进行分析，以识别视频中的人物行为模式。

[0003] 行为识别是计算机视觉与模式识别中一个重要而又有难度的课题。它在许多方面有着广泛的应用前景，比如智能监控、人机交互、虚拟现实、智能安防等。当今社会随着经济的快速发展，安全问题越来越受到人们的关注，越来越多的场所都安装了视频监控摄像头，每天都有大量的监控视频产生，现在人们普遍采取人工进行监控，这就需要投入大量的人力、物力、财力；除此之外，人在长时间的视频监控时，注意力会下降，导致一些紧急情况无法被有效检测到，耽误了救援等工作的时间，造成大量的损失。为了解决这个问题，如果可以通过计算机自动视频监控，那么将有效减少这方面的投入，并能够得到及时的警报信息，以方便采取措施。

[0004] 行为识别的一般方法是，首先提取视频帧的特征，然后利用分类模型进行识别。其中，从视频中提取有效的特征表达很重要，直接影响整个系统的性能。

[0005] 传统的行为识别方法，使用不同的手工设计特征。深度学习方法则利用神经网络学习到的特征，由于深度网络的非线性，它可以包含更多的内在信息，可以极大提升特征表达与泛化能力。

[0006] 传统的行为特征提取方法主要有：

1)基于低层特征如运动目标检测、跟踪的方法。该方法是提取低层图像特征进行分析，方法比较简单直观。可以使用的低层特征主要包括前景目标的运动速度、方向、光流、目标形状轮廓、运动轨迹等。这类特征可以忽略静止的无关信息，重点关注运动目标，可以减少视频帧中无关信息的干扰。另外，该方法提取特征也相对简单，但其中也存在一些严重的问题。比如，这类特征严重依赖目标跟踪算法，如果跟踪效果太差，将会对最后的结果造成极大的影响。而现实世界中的视频，往往又包含各种干扰，比如杂乱的背景，其它运动目标。这都导致这类方法在进行日常视频行为分析效果较差，所以这类特征往往鲁棒性较差，很难应用到实际的场景中。

[0007] 2)基于时空描述子特征的方法。这类特征利用手工特征描述的行为模式。比如时空兴趣点方法，它用一些不关联的点对人体动作进行描述。除此之外，还有密集轨迹(Dense Trajectories)，改进轨迹(Improved Trajectories)，尺度不变特征变换(SIFT)，梯度直方图(HOG)，光流直方图(HOF)等方法，这些手动设计的特征比单纯的运动特征更鲁棒，但局部描述子的计算量较大，也容易受到噪声的干扰。目前这些特征描述取得了不错的效果，但其提升的空间已经很小，需要寻找一个更强大有效的特征。

[0008] 3)基于中层语义理解的特征。这类信息一般使用一个统一的人体模型，这个模型

将人体划分为不同的部分，如头、肩、手臂、腿等，这种表示可以取得较高的精度，也相对比较鲁棒，但模型的构建相对复杂，需要大量的工作。为了提升准确率，许多文章利用RGB-D深度图进行检测。

[0009] 以上三种是传统行为识别中主要用到的特征。林贤明为了提升行为识别的性能，采用多个特征进行有效融合，取得了一定的成果。Lai等人把每帧作为一个实例，使用多实例学习方法，系统可以同时推理帧标签和视频标签。Haoi等人使用结构化SVM去建模视频帧之间的关系，并在观察部分事件时预测整个事件。

[0010] 传统学习方法在一定程度上取得了不错的结果。但随着深度学习的出现，许多领域的性能都得到了极大的提升。深度学习模拟大脑的工作机制，它采用了多层的网络结构，它是一种非线性的模型，具有强大的数据拟合能力和学习能力，通过对视觉对象进行抽象，可以无监督的从数据中学习到内在信息。通过深度学习得到的特征，与人类的感知结果相似，往往包含了一定的语义信息，更利于对视觉对象进行分析识别。

[0011] 使用卷积神经网络进行静态图片分类已经取得极大的成功。但是因为视频的复杂结构和噪声干扰，行为识别的关注度较少。如果我们把每帧视频帧当作静态图片进行处理，然后对视频所有帧的结果求平均得出这个视频所属分类。但是这样的方法仅依赖静态帧丢失了太多的时间相关信息，这样会导致性能下降。

[0012] 就行为识别而言，主要存在两个问题。

[0013] 1、提取特征与视频语义之间的鸿沟。人类手工设计的特征，在过去中取得了有效的成果，但现在的手工特征表达已经到了一个瓶颈，很难再取得大的提升，特别是处理视频这种复杂问题，视频的背景干扰，帧率变化，光照变化，视角变化和摄像机的运动等等问题都严重影响系统的性能，所以需要一个更好的特征描述。

[0014] 2、另一个值得关注的问题就是如何建模视频的时间信息。仅使用空间特征，无法准确描述视频。为了解决这个问题，必须考虑加入时间信息。不同的动作拥有不同的序列，将时域信息加入可以有效地提升识别效果。

发明内容

[0015] 本发明的发明目的是提供一种用于视频的基于双向长短时记忆单元的行为识别方法，通过采用多层深度学习的特征代替传统的人工特征，提高分类的准确率，通过采用双向长短时记忆(Bi-directional Long Short-Term Memory)捕获时间信息，获得更多的时域结构信息，以提高行为识别的效果。

[0016] 为达到上述发明目的，本发明采用的技术方案是：一种用于视频的基于双向长短时记忆单元的行为识别方法，包括以下步骤：

(1)输入视频序列，提取所述视频序列中的RBG帧序列和光流图像；

(2)训练深度卷积网络：分别训练RGB图像深度卷积网络和光流图像深度卷积网络；

(3) 提取网络的多层特征，其中至少提取第三卷积层、第五卷积层、第七全连接层的特征；对全连接层得到一个固定大小的向量；对卷积层特征进行和池化，加入时间信息；

(4) 利用从卷积神经网络得到的各层特征向量，对采用双向长短时记忆单元构建的递归神经网络进行训练，并输入测试集利用Softmax得到视频中光流帧每帧的概率矩阵和RGB帧每帧的概率矩阵；

(5) 分别对视频中所有光流帧的概率矩阵取平均，对视频中所有RGB帧的概率矩阵取平均，最后融合光流帧和RGB帧的概率矩阵，取概率最大的类作为最后的分类结果，由此实现行为识别。

[0017] 优选的技术方案，步骤(2)中，选择第三卷积层、第五卷积层和第七全连接层作为特征表达。

[0018] 所述和池化为：

在时间t第m层特征图是F_m^t，使用下面的公式进行池化：

$$Desc = \sum_{j=1}^{N} F_m^j(x,y)$$

其中，j是池化帧，N是时域池化范围，最后描述子的维度是$H_t \times W_t \times C \times T$，$H_t$是特征图的高度，$W_t$是特征图的宽度，C是特征图通道个数，T是视频的总帧数。

[0019] 进一步的技术方案，在所述和池化后步骤(4)之前，使用空间金字塔最大池化，使用两层金字塔，得到维度为定长的特征向量。

[0020] 优先地，N为15。

[0021] 由于上述技术方案运用，本发明与现有技术相比具有下列优点：

1、本发明采用多层深度学习的特征代替传统的人工特征，不同层的深度特征表征了不同的信息，多层特征的组合可以提高分类的准确率。

[0022] 2、本发明采用双向长短时记忆(Bi-directional Long Short-Term Memory)捕获时间信息。双向长短时记忆通常被用来处理序列问题，单向的长短时记忆单元只能对单方向时间进行建模，但有时当前帧不仅要参考前一帧，还要考虑后一帧的信息，这样，使用双向长短时记忆单元便可以对两个方向进行建模，保证了行为识别的准确性。

附图说明

[0023] 图1是本发明实施例的方法框架构图；

图2是对相同的输入图片不同层的特征图的可视化；

图3是LSTM架构示意图；

图4是本发明实施例中的双向LSTM架构示意图；

图5是UCF101数据集示意图；

图6是不同层的可视化示意图。

具体实施方式

[0024] 下面结合附图及实施例对本发明作进一步描述：

实施例一：参见图1所示，为一种用于视频的基于双向长短时记忆单元的行为识别方法，使用深度学习到的特征结合双向长短时记忆单元进行行为识别。为了选择一个强有力的特征表达，使用多层的深度学习特征替换传统的手工设计特征，提高行为识别的性能。为了充分发掘时序信息，使用双向长短时记忆单元(Bi-LSTM)进行建模，它可以捕获两个方向上时序序列的变化，提供的信息要优于单向的长短时记忆单元。

[0025] 1、卷积神经网络及其多层特征

为了提取有效的表达特征，需要训练一个深度卷积网络。本实施例使用简单有效的Caffe架构搭建系统。使用ImageNet数据集对模型进行预训练。ImageNet数据集包含大量的图片，这可以保证模型的泛化能力。之后，迁移预训练好的模型到实际数据集上进行微调。使用类似双流的网络架构。预先提取RGB帧和视频序列的光流并保存到本地磁盘。本实施例使用224×224大小的帧作为输入。

[0026] 不同层包含不同的信息。网络的前几层有更多的低层特征，比如边缘信息。网络的后几层会更加抽象，编码了视频更多的语义信息。图2是对相同的输入图片不同层的特征图的可视化。从图中，我们发现不同卷积层包含不同的信息表示。它们对同一幅图片有不同的响应。我们通过实验最后选择第三卷积层(conv3)，第五卷积层(conv5)和第七全连接层(fc7)作为我们的特征表达。第一卷积层(conv1)和第二卷积层(conv2)的泛化能力太差，对最后的识别结果作用不大。

[0027] 卷积神经网络各层的参数如下：

层	Conv1	Pool1	Conv2	Pool2	Conv3	Conv4	Conv5	Pool5	Full6	Full7	Full8
卷积核	7×7	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
步长	2	2	2	2	1	1	1	2	-	-	-
通道数	96	96	256	256	512	512	512	512	4096	2048	101

在提取特征图之后，对特征图使用和池化(Sum Pooling)。假如在时间t第m层特征图是F_m^t，使用下面的公式进行池化：

$$Desc = \sum_{j=1}^{N} F_m^j(x,y)$$

其中，j是池化帧，N是时域池化范围。这里选择15帧。最后描述子的维度是$H_f \times W_f \times C \times T$。$H_f$是特征图的高度，$W_f$是特征图的宽度，C是特征图通道个数，T是视频的总帧数。对于全连接层，直接使用网络输出的4096维特征向量。

[0028] 2、递归神经网络

现有技术中主要有两种类型的神经网络，前馈神经网络和递归神经网络。前馈神经网络在许多应用中已经取得了成功，但是它却无法很好处理序列问题。相反，递归神经网络的特点使之对时域进行建模很简单。递归神经网络以序列作为输入。对视频序列来说，输出与当前帧和前一帧有关。假如给定的输入序列表示为$x=\{x_1, x_2, \ldots, x_T\}$，有如下公式：

$$h_t = f(W_{ih}x_t + W_{hh}h_{t-1} + b_h)$$

其中h_t表示在时间t时隐藏层的激活值，W_{ih}表示输入层到隐藏层的权重矩阵，W_{hh}表示隐藏层之间的权重矩阵，b_h是偏移，f是激活函数。最后，通过如下公式得到输出：

$$y_t = W_{ho}h_t + b_o$$

其中，W_{ho}表示隐藏层到输出层的权重矩阵，b_o是输出偏移。

[0029] RNN主要问题是它只能对短时间的序列进行有效建模，随着网络深度加深，会导致梯度弥散，以前的帧对当前的作用效果太小。为了解决该问题，长短时记忆单元(LSTM)引入三个门去保持网络状态。三个门分别是输入门，遗忘门和输出门。输入门控制以前通过加入这三个门，它使在处理序列问题中，梯度变化更加平稳。LSTM使用一个比传统递归神经网络复杂的架构去提高处理长时序列的性能。图3描述了LSTM架构。

[0030] 3、双向长短时记忆单元

虽然LSTM能够捕获长时序列信息，但它只是单向的。也就是说，LSTM中当前帧只受之前帧影响。本发明中，为了强化这种关系，使之扩展成双向的。即，在处理当前帧时，还要考虑之后帧的影响。其实现方法是，再加上一层从后向前处理。本实施例的双向LSTM模型如图4所示。第一层是从前向后处理序列，第二层是从后向前处理序列。最后，这两层的结果会共同作用输入到Softmax分类器中。单向LSTM只能捕获一个方向的时间演变信息，但双向LSTM可以对双向的时序结构进行建模，因此它能够捕获更多的时域结构信息。

[0031] 在UCF101和HMDB51数据集上测试该模型。UCF101数据集包含13,320个视频序列，共101个行为类别。如图5所示，这些视频都不是由专业人员拍摄的。它分为三部分，每个部分分别划分训练集和测试集。在每个部分上测试本实施例的方法并将结果取平均值。HMDB51数据集来自多个方面，比如电影或者网络视频。这个数据集由6766个视频序列组成，共51个行为类别。它比其它数据集更具挑战性，因为它的背景环境更复杂。数据集也被划分为三个部分。取三个部分结果的均值作为最后的结果。

[0032] 为了选取有用的CNN层，在UCF101数据集上比较不同层的实验结果。从图6中可以发现，对于一幅图片输入，不同层有不同的激活。前几层有更多的细节信息，后几层更加抽象。但是也发现第一卷积层和第二卷积层包含过多的干扰信息。

[0033] 下一步进行定量分析。为了简单快速比较，每个视频只采样10帧。短视频一方面特征已经足够具有代表性，另一方面减少了实验时间。表1中的结果是在空间网络上以RGB帧作为输入产生的结果。从表1中，也能发现最早的几层效果不是最好，这会影响最后的结果。

[0034] 表1 不同单层在UCF101上性能

层	conv1	conv2	conv3	conv4	conv5	fc6	fc7
准确率	33.5%	44.7%	68.2%	68.9%	69.1%	60.3%	61.4%

继续不同层组合的结果。使用多层组合可以取得比单层更好的结果。这证明了多层的CNN特征更加强大。如表2，在实验中发现使用第三卷积层(conv3)、第五卷积层(conv5)和第七全连接层(fc7)取得最好的结果。

[0035] 表2 不同组合层在UCF101上的性能

层数	准确率
conv3 + conv4	70.1%
conv4 + conv5	69.2%
conv3 + conv5	70.5%
conv3 + conv5 + fc6	70.3%
conv3 + conv5 + fc7	70.8%

再对本实施例的双向长短时记忆单元(Bi-LSTM)和平均模型和单向长短时单元进行对比试验。平均模型直接平均每个Softmax的得分去获取最后的结果。使用单层LSTM在单层LSTM模型中。所有三个模型已经提出fc7作为输入。从表3中的结果发现双向LSTM优于平均模型和单向LSTM。

[0036] 表3 UCF101上不同模型性能

模型	准确率
平均模型	61.4%

单向LSTM	62.1%
双向LSTM	63.5%

最后，表4描述了各种方法的准确率。证明加入了多层CNN特征和双向LSTM的方法可以有效地提高行为识别性能。

[0037] 表4 不同方法在数据集UCF101和HMDB51上的结果

模型	UCF101	HMDB51
STIP+BovW (2011, 2012)	43.9%	23.0%
Motionlets (2013)	-	42.1%
DT+MVSV(2014)	83.5%	55.9%
iDT+FV (2013)	85.9%	57.2%
iDT+HSV (2014)	87.9%	61.1%
Two-Stream (2014)	88.0%	59.4%
LRCN (2015)	82.9%	-
BSS (2015)	88.6%	-
Composite LSTM (2015)	84.3%	-
本实施例	88.9%	62.3%

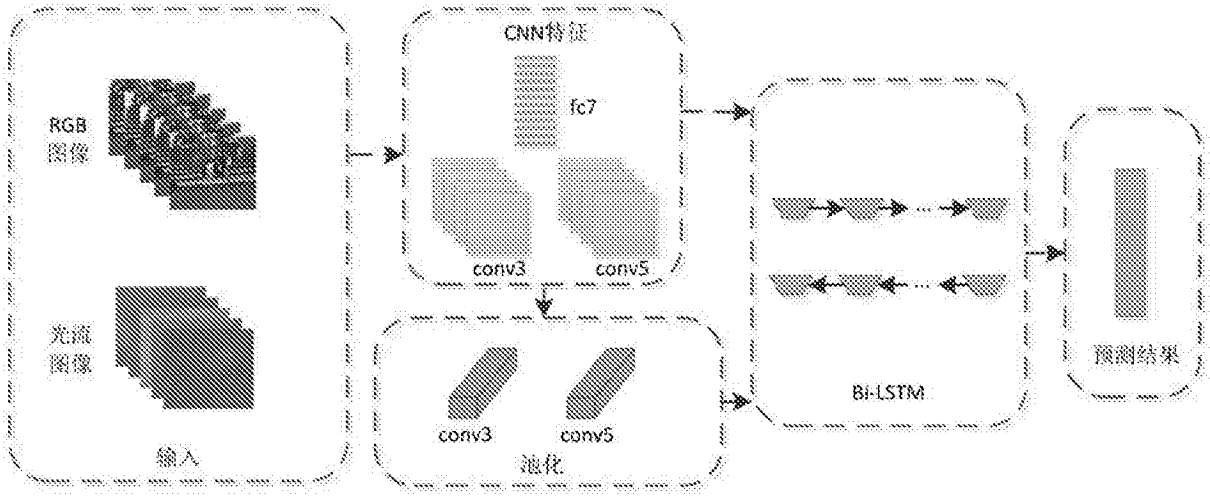

RGB
图像
光流
图像
输入
CNN特征
fc7
conv3
conv5
conv3
conv5
池化
BI-LSTM
预测结果

图1

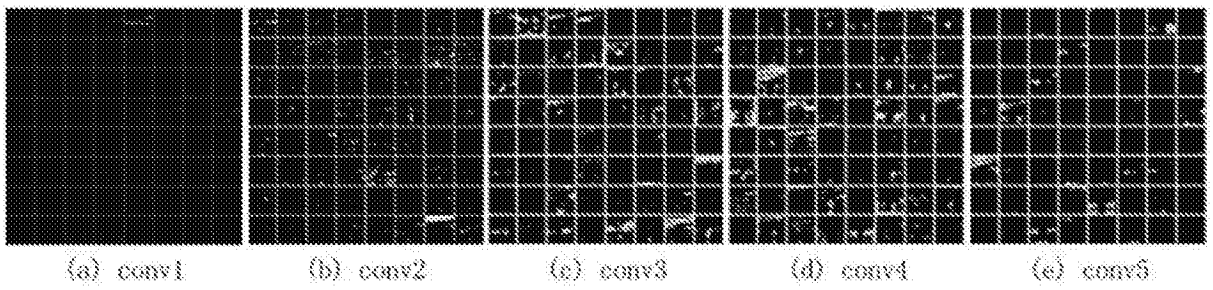

(a) conv1
(b) conv2
(c) conv3
(d) conv4
(e) conv5

图2

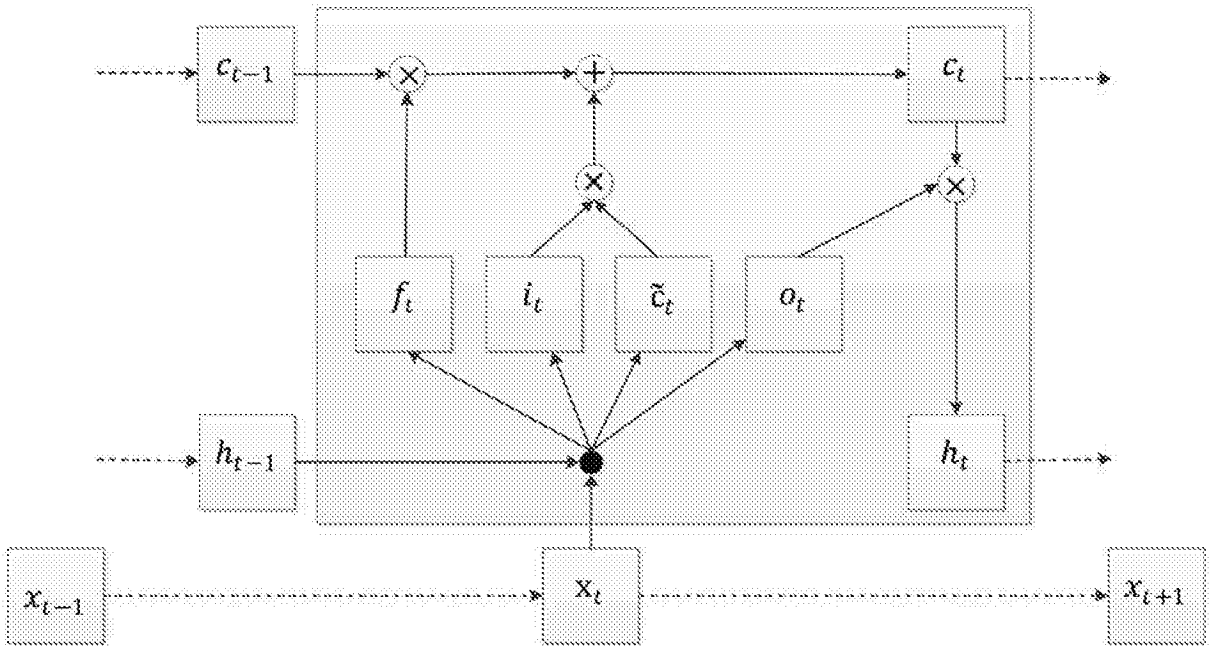

c_{t-1}
c_t
f_t
i_t
$\tilde{c}_t$
o_t
h_{t-1}
h_t
x_{t-1}
x_t
x_{t+1}

图3

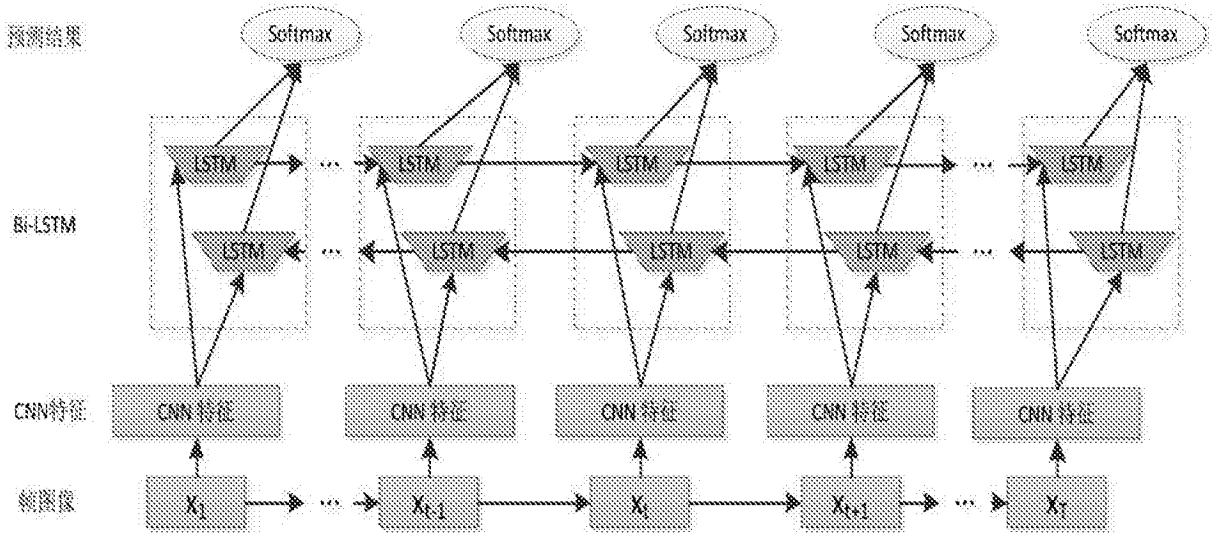

预测结果
Softmax
Softmax
Softmax
Softmax
Softmax
Bi-LSTM
LSTM
CNN特征
CNN 特征
帧图像
X_1
X_{t-1}
X_t
X_{t+1}
X_T

图4

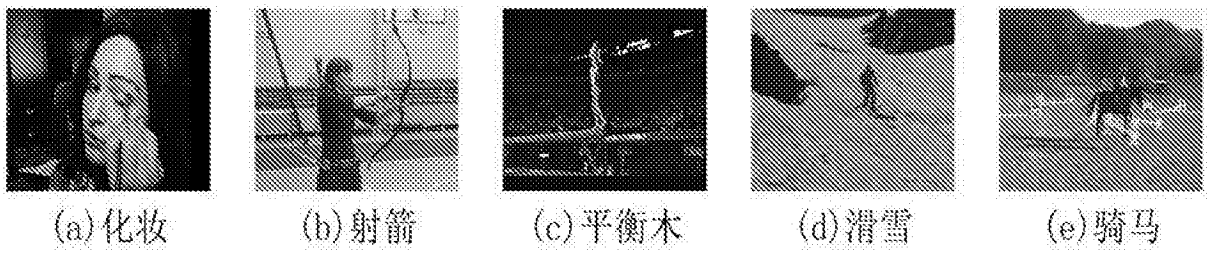

(a)化妆
(b)射箭
(c)平衡木
(d)滑雪
(e)骑马

图5

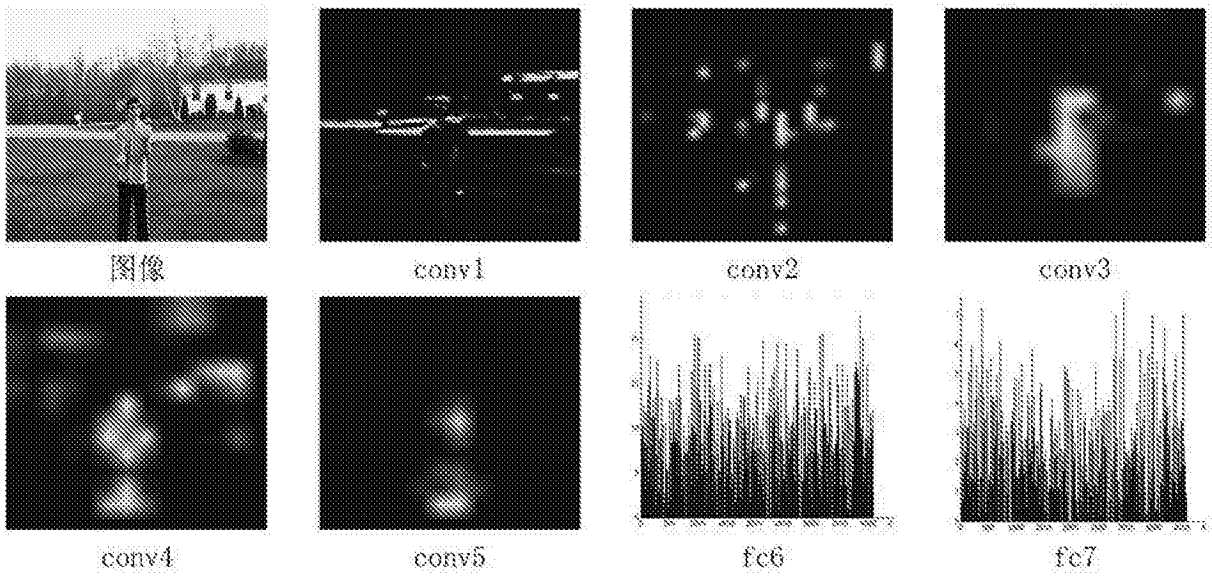

图像
conv1
conv2
conv3
conv4
conv5
fc6
fc7

图6