基于视频场景深度学习的人物语义识别模型

高 翔[1],陈 志[1],岳文静[2],龚 凯[1]
(1. 南京邮电大学 计算机学院,江苏 南京 210023;
2. 南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:为有效分析和整合与人物行为相关的视频语义线索,提出一种基于视频场景深度学习的人物语义识别模型。该模型由中层语义特征提取、多通道语义特征融合、整体精调和语义识别等组成。首先实现底层图像到中层特征抽取,利用卷积神经网络算法并行获取视频场景关键帧集中的人物身份、人物行为、上下文环境等通道语义;再将中层特征融合到同一个语义融合层,通过多层语义卷积神经网络来整合上述语义,使用损失函数来学习不同通道语义之间的潜在关系,提高人物语义融合的鲁棒性;最终通过大间隔的损失函数来精调整个网络的参数,利用 SVM 分类器完成视频人物语义识别。实验结果表明,该模型在特定的数据集上具有较高的准确率,能够高效地识别视频人物语义。

关键词:视频挖掘;深度学习;卷积神经网络;人物语义;支持向量机
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2018)06-0053-06
doi:10.3969/j.issn.1673-629X.2018.06.012

Human Semantic Recognition Model Based on Video Scene Deep Learning

GAO Xiang[1], CHEN Zhi[1], YUE Wen-jing[2], GONG Kai[1]
(1. School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to effectively analyze and integrate the video semantic clues related to human behavior, we propose a human semantic recognition model based on the video scene deep learning. The model consists of middle layer semantic feature extraction, multi-channel semantic feature fusion, global fine-tuning and semantic recognition. Firstly, it achieves the extraction of low-level image to the middle layer feature, and uses the convolutional neural network algorithm to concurrently extract the channel semantics of human identity, human behavior, and context in video scene key frame set. Then it fuses the middle layer features to the same semantic fusion layer, integrating those semantics through the multilayer semantic convolutional neural network, and uses the loss function to learn the potential relationship among the different semantic channels, so as to improve the robustness of human semantic fusion. Finally it fine-tunes the whole network parameters through the large interval loss function, and uses SVM classifier to complete video human semantic recognition. Experiments show that the proposed model has a high-accuracy rate on the specific data set, and can effectively recognize the video human semantic.

Key words: video mining; deep learning; convolution neural network; human semantics; SVM

0 引 言

视频语义是对视频信息所包含事物的状态描述和逻辑表示,涉及人和物的动作、表情、音频、图像序列等信息[1-2]。视频语义分析与识别是对视频包含的语义信息进行特征提取、整理、分析与识别的过程,涉及人的视觉机理、图像识别、机器学习、模式识别和深度学习等领域[3]。

在对视频中有序的帧图像进行语义分析中,由于

收稿日期:2017-04-08 **修回日期**:2017-08-16 **网络出版时间**:2018-02-07
基金项目:国家自然科学基金(61501253);江苏省基础研究计划(自然科学基金)项目(BK20151506);江苏省“六大人才高峰”第十一批高层次人才选拔培养资助项目(XXRJ-009);江苏省重点研发计划(社会发展)项目(BE2016778);江苏省 2016 年度普通高校研究生实践创新计划项目(SJLX16_0327);南京邮电大学科研项目(NY217054)
作者简介:高 翔(1991-),男,硕士研究生,研究方向为数据挖掘;陈 志,副教授,通信作者,CCF 会员(14587M),研究方向为移动物联网、无线传感器网络、数据挖掘。

一段视频中可能包含多个场景,而这些场景又由一组有序的帧图像组成,为了更好地分析视频语义,需要对视频进行预处理,包括把视频中的内容按某种方式进行镜头分割并场景化[4-5]。在上述视频人物语义分析中,首先将通过镜头检测和寻找镜头变化的方法对视频进行分割,其次将找出镜头中的关键帧集,并通过计算所有镜头的关键帧图像之间的相似度来进行聚类,最后研究视频场景化中的人物语义[6-7]。

视频人物语义分析往往是以研究视频中的人物行为语义为中心,同时辅助视频中除人物以外的事物所构成的上下文环境对象的语义,来提高分析人物语义信息的准确性[1]。目前视频语义分析一般都是通过学习图像特征这种方法,图像特征主要包括低层特征和中层特征。低层特征是基于视频的像素经由各种变换而来的,没有具体的语义含义。对于简单行为的识别,低层特征具有很好的描述效果,但通常难以对真实场景下的复杂行为进行有效建模[6-7]。

文中提出一种人物语义识别模型(DVSM),该模型由语义通道层和语义融合层构成,从人物身份、人物行为、上下文环境等通道对视频预处理好的场景图像运用卷积神经网络进行处理,从底层图像抽取中层特征,再将这些中层特征融合到语义融合层来识别视频人物语义。

1 相关工作

1.1 视频场景划分

镜头分割是视频场景预处理的第一步,现如今比较成熟的镜头分割方法有 X2 直方图匹配算法与梯度法。基于 X2 直方图匹配与梯度法的镜头检测算法检测视频中的镜头切换和淡入淡出。该算法是通过计算视频中连续两帧图像的直方图差值来检测镜头切换。除切换外,另一个重要的镜头连接方式是淡入淡出,其特点是视频帧的画面先渐渐暗下去,然后再亮起来,因此每帧画面的相邻像素相关性都会先变小再变大,而每两个像素的梯度恰好能代表他们的相关性。

关键帧提取是要获取视频场景中能够代表镜头内容的图像。Li 等提出一种基于非相邻帧比较的关键帧提取算法[1]。该算法的思想是选择镜头中的第一帧作为第一个关键帧和参考帧,然后计算后续帧和当前参考帧的差异,当差异大于预定的阈值时,则选后续帧为关键帧和参考帧,重复上述过程直到镜头结尾。

镜头聚类是完成视频场景预处理的重要步骤,首先通过 HSV 空间中的颜色直方图来描述关键帧的整体颜色特征,并以此作为特征值进行关键帧聚类;接着通过计算关键帧之间的相似度值作为输入来计算镜头相似度以对上述颜色直方图特征进行匹配;最后计算簇中元素间的最大相似度,当相似度值大于一个预先设定的阈值时,将这两个簇合并为一个簇,直到簇间距离都小于阈值则聚类终止。归为一类的镜头集,即为场景,聚类结束即完成对将视频的场景划分。

1.2 视频语义分析

视频中人物的语义信息具体可以细分为人物的身份信息、动作、表情、语音等几个主要方面。现有融合语义主题的方法将每幅图像的视觉特征表示为一个视觉“词袋”,设计一个概率模型分别从视觉模态和文本模态中捕获潜在语义主题,采用一种自适应的不对称学习方法融合两种语义主题[8]。Atan 等提出了基于多用户和多处理的系统学习框架来识别视频中的人脸[9],与已有的强化学习技术相比,在高度动态的环境中,这种方法学习接近最佳状态的收敛速度更快。Kumar 等提出了一种新颖方法来挖掘新闻视频语义[10],首先通过基于人脸识别来命名新闻视频中人物,并对视频中人物聚类成多个社区,其次再通过语义分析模型分析出社区之间的联系。Liang 等提出了一个表达深度模型来自然地融合人和周围的环境以高层次地在静止图像中理解动作[11]。特别地,训练了一个深度置信网络以从不同的噪声源中融合信息。Zhan 等提出了一种基于稀疏表示的核判别分析加 KNN 的视频语义方法[12],通过引入核分类功能到 KSVD 字典优化算法来建立可判别模型,通过该模型完成稀疏表示特征到高位空间的映射,使用基于优化的稀疏表示的加权 KNN 方法来分析视频语义。

Zhang 等提出了一个深层次的学习策略,以融合多复杂事件识别的语义线索[13]。通过回答如何共同分析人类行为、对象和场景来解决识别任务。首先,每种类型的语义特征被馈送到一个相应的多层特征抽象的路径,由一个融合层连接所有不同途径。然后,通过无人监督的跨通道编码方式学习语义线索相互作用的关联性。最后,通过微调架构上大幅度的目标,来回答语义线索如何组成一个复杂的事件。相比于传统的特征融合方法,该方法有效地融合了识别的水平特征[12],但该语义线索局限在人类行为、对象和场景等方面,缺乏对每一个人物的身份信息提取和分析;此外,该方法在自编码学习过程产生的参数数量太多,增加了深度学习的难度。

文中在改进上述视频语义模型的基础上,利用视频场景深度学习构建人物语义识别模型。

2 基于深度学习的视频人物语义识别模型

2.1 视频人物语义识别模型框架设计

图 1 给出了基于深度学习的视频人物语义识别模型(DVSM)框架。该框架包括三个部分:中层语义特

征提取、多通道语义特征融合、整体精调和语义识别。在该框架中,卷积网络的输入层把所有场景中的关键帧图像集作为提取层的输入,分别提取图像集中的人物语义 $X_P^{(0)}$ 、行为语义 $X_A^{(0)}$ 、上下文环境语义 $X_S^{(0)}$ 三个通道的低层特征,并行地通过卷积神经网络学习降低每一个通道语义的低层特征向量维数,完成三个通道的中层语义特征提取的训练。每一个通道语义特征提取过程包括卷积、子采样过程和全连接,如图 2 所示。在视频人物语义识别中,将中层语义特征作为多融合层的输入 $Z^{(l)}=[X_P^{(l)},X_A^{(l)},X_S^{(l)}]$,通过多通道卷积神经网络学习方法来完成多通道语义特征融合 $Z^{(l+1)}=[X_P^{(l+1)},X_A^{(l+1)},X_S^{(l+1)}]$,其中每一层的语义特征之间的关联关系通过引入一个损失函数来学习调整参数,最终把多通道语义特征融合的结果作为识别层的输入,通过运用大间隔的损失函数精调整个网络学习的参数,SVM 分类器完成识别人物语义的任务。

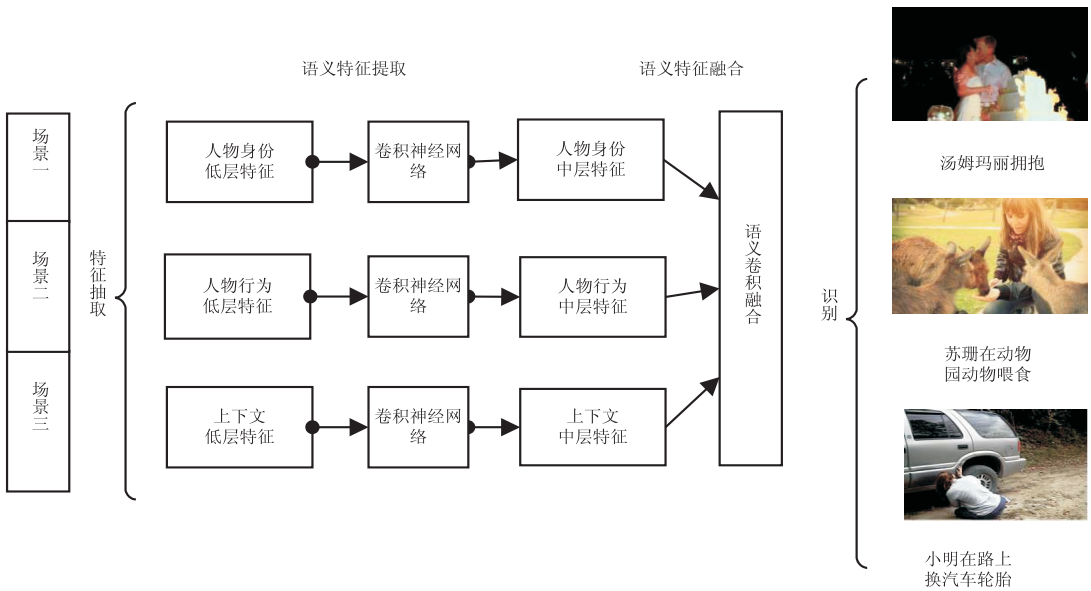

图 1　基于视频场景深度学习的人物语义识别模型框架

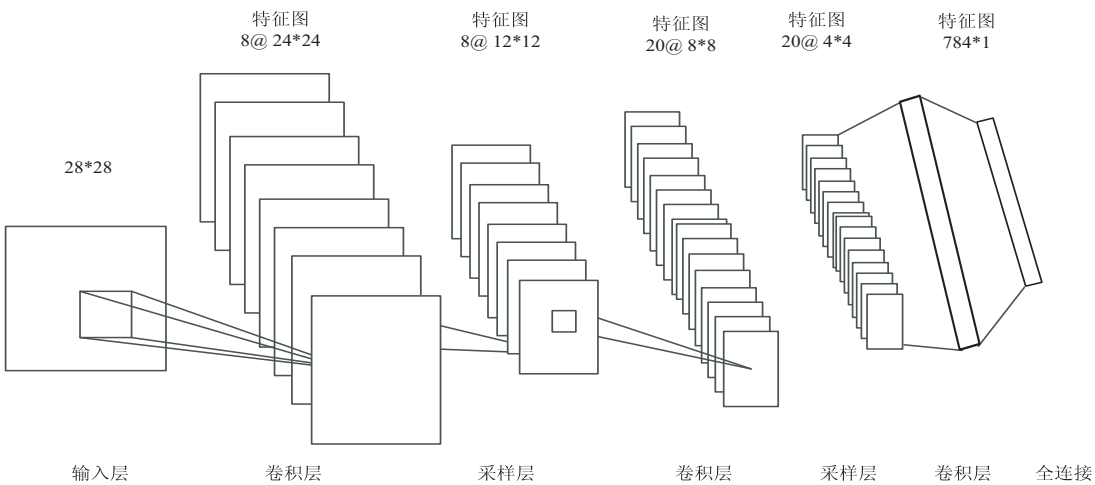

图 2　通道语义特征提取过程

2.2　通道中层语义特征提取

通道中层语义特征提取主要是卷积神经网络中的卷积、采样和全连接过程。卷积本质上是通过一个或多个可训练的滤波器即卷积核,对原特征向量做一次或多次非线性变化。为了更好地描述每两层之间的卷积过程,通过 (N_l,b_l*b_l) 来描述第 L 层神经元;通过多个可训练的滤波器 $f(n*n)$ 向量和多个连接表 N_l*N_{l-1} 来描述 L 层和 $L-1$ 层之间神经元的卷积运算。通过多个可训练的滤波器 $f(n*n)$ 向量卷积一个输入为 $m*n$ 维的图像,然后加上偏置 b ,得到卷积层的输出特征图,用 (N_l,b_l*b_l) 描述,其中 N_l 表示第 L 层的特征图个数, b_l 表示第 L 层的特征图维数。第一层输入的是图像,后面阶段输入的是从前一层抽取的卷积特征图集合的一个子集。具体要几个特征图来卷积构成后一层的一个特征图,需要先设定好一张两层特征图之间的连接表,该表记录着两层特征图之间的连接关系。

以行为语义通道为例,卷积层公式如下:

$$X_{A_{ij}k}^{(l)} = \sigma\left(\sum_{k\in D^l}\left(\sum_{i=1}^{n}\sum_{i=1}^{n} F_{A_{ij}}^{(l)} * X_{A_{ij}k}^{(l-1)} + b^l\right)\right) \quad (1)$$

其中，$F_{A_{ij}}^{(l)}$ 表示行为语义通道上的第 L 层卷积核；D^l 表示第 $L-1$ 层需要输入的特征图集合；b 表示卷积操作后需要加上的偏置。

子采样本质上是给卷积层中得到的每一个特征图进行降维。典型的操作一般是对输入图像中大小为 $n*n$ 块的所有像素进行求和，这样输出图像在两个维度上缩小了 n 倍。文中将每一幅特征图中每个不重复邻域的两个像素求和，变为一个像素，然后通过乘性偏置 $\beta x+1$ 加权，再增加加法偏置 $bx+1$，然后通过 sigmoid 激活函数产生一个缩小二倍的特征映射图 $Sx+1$。这里以行为语义通道为例，卷积层公式和采样层公式如下：

$$X_{A_{ij}}^{(l)} = f(\beta_{A_{ij}}^{(l)} * \mathrm{down}(X_{A_{ij}}^{(l)}) + b_{A_{ij}}^{(l)}) \quad (2)$$

其中，down 函数表示子采样函数。每个输出特征都对应一个乘性偏置 β 和一个加性偏置 b。

全连接是将卷积核在前一层所有的特征图上做卷积操作，将特征向量降为 $1*n$ 维的向量。文中将每个通道上的语义通过各自全连接层，输出一个 $1*n$ 向量特征。

2.3 多通道语义特征融合

在完成对人物身份、行为、上下文环境的中层语义特征抽取的基础上，把上述中层语义特征作为多通道语义融合层的输入，构造向量矩阵 $\boldsymbol{Z}^{(l)} = [X_{\mathrm{P}}^{(l)}, X_{\mathrm{A}}^{(l)}, X_{\mathrm{S}}^{(l)}]$ 作为多通道语义融合的输入，然后通过多语义卷积神经网络学习方法来完成多通道语义特征的完全提取，公式如下：

$$\boldsymbol{Z}^{(l+1)} = \sigma(F^{(l)}\boldsymbol{Z}^{(l)} + b^{(l+1)}) \quad (3)$$

其中，$\boldsymbol{Z}^{(l+1)}$ 表示融合层三层中层语义的卷积输出。但是由于视频中存在语义噪声，会造成语义抽取的不完整或者丢失，为了让文中提出的语义模型可以学习到多通道语义之间的关联关系，增强语义融合的鲁棒性，定义式 5 作为融合语义的损失函数。

$$\boldsymbol{L}_{\mathrm{pas}} = \| \boldsymbol{Z}^{(l)} - \boldsymbol{Z}_{\mathrm{pas}}^{(l)} \|^2 \quad (4)$$

其中，L_{pas} 表示人物身份语义特征舍弃，只融合动作行为和上下文环境语义，即 $Z_{\mathrm{pas}}^{(l)} = [0, X_{\mathrm{A}}^{(l)}, X_{\mathrm{S}}^{(l)}]$。

语义融合的完整损失函数如下：

$$L = L_{\mathrm{pas}} + L_{\mathrm{pas}} + L_{\mathrm{pas}} + L_{\mathrm{pas}} \quad (5)$$

2.4 整体精调与语义识别

通过有监督的机器学习来调整整个网络所有层参数并完成语义识别任务，特别是在 SVM 分类器中加入最大间隔分类来构造损失函数。一种流行的方法是训练多个一对多的模型，一个类别对应一个模型，其中每个模型计算真实类别 $y \in \{1, -1\}$ 和预测类别之间的损失，然后将融合层特征向量 $\boldsymbol{Z}$ 作为前向传播的训练数据，W 作为融合层和识别层之间的权重参数，大间隔损失函数如下：

$$\max(1 - W^{\mathrm{T}} * z * y, 0) \quad (6)$$

为将式 6 加入到深度学习网络中，借鉴 Zhang 等在多层语义融合时运用的 l_2 - loss 函数，考虑到该函数的权值衰减问题，最终融合层的大间隔代价函数类似于二类 SVM 分类器公式[13]。

$$L(W) = \min_W \frac{1}{2} * W^{\mathrm{T}} * W + C\sum_{n=1}^{N} \max(1 - W^{\mathrm{T}} * z * y, 0) \quad (7)$$

为了简化多层框架的训练过程，将上述二类扩展到多类，与之相匹配的 l_2 - loss 函数如下[13]：

$$\sum_{k\neq Y}^{K}\sum_{y\in Y} \max(1 - W_y^{\mathrm{T}} * z + W_k^{\mathrm{T}} * z, 0)^2 \quad (8)$$

其中，Y 表示样本的类别集合；W_k^{T} 表示语义 K 与融合层的连接权重。

3 实验与结果分析

3.1 数据集和预处理

选择 OA 视频集中的事件作为实验数据，该数据集是发生在办公室里面人物的日常行为，是公开的 RGB-D 视频数据集，包含 1 180 个视频序列，10 个以上人物，两个办公室地点，每一个事件同一个人做两次，还包括两个人物之间的交互事件。上述数据集分成两组子数据集：OA_1 和 OA_2，每一个子数据集有 10 类事件，OA_1 是单个人物的事件，OA_2 是两个人物的交互事件，具体如表 1 所示。

表 1 OA 视频场景数据集

Seq	OA_1	OA_2
1	A answer the phone in office$_1$	B ask C and away in office$_2$
2	A arrange in office$_1$	B called C away in office$_2$
3	A eat in office$_1$	B and C carrysth in office$_2$
4	A go in office1 to takesth and go out	B chat with C in office$_2$
5	A go to work in office$_1$	C deliversth to B in office$_2$
6	A look for in office$_1$	B eat and chat with C in office$_2$
7	A wiping in office$_1$	Having-guest in office$_2$
8	A sleep in office$_1$	C Seek-help B in office$_2$
9	A take water in office$_1$	Shaking-hands in office$_2$
10	A wander in office$_1$	B showingsth to C in office$_2$

实验数据集的预处理主要是将视频文件转换成文本文件格式数据。首先通过对视频进行场景分割和聚类，每个视频由一系列关键帧组成的场景集合表示，聚类好的每一个视频场景需要指定相应类别；然后对每个场景中的图片分别进行人脸、动作和上下文环境检测与特征提取，生成对应的人脸、动作和上下文环境的

三张图片;最后通过对上面检测出来的三张图片分别进行灰度化与二值化,重新统一图片大小为 28 * 28,将图片的所有像素按行遍历输入到文本文件中的一行大小为 1 * 784,并在末尾加上所属类别。该文本文件就是三通道语义中层特征提取的训练数据集,具体包括三个训练集:person_train. txt、action_train. txt、context_train. txt;三个测试集:person_test. txt、action_test. txt、context_test. txt。

3.2　中层特征提取训练过程

根据第 2 节中提出的语义识别模型进行实验,其中中层特征提取包括人物身份、人物行为、人物所处的上下文环境的中层特征,三个通道并行利用 6 层卷积神经网络来训练 3.1 节预处理出的训练集。主要分成以下几步:

(1)卷积网络初始化。

实验的初始化主要是对卷积网络初始化卷积层和输出层的卷积核和偏置,其中卷积核和权重进行随机初始化,而对偏置进行全 0 初始化。

(2)前向传输计算。

实验的卷积网络按照输入层、卷积、采样、输出层来构成。实验中的每一个卷积层的卷积核大小为 5 * 5,采样层的采样规模为 2 * 2。实验经过多 2 层卷积 2 层采样最终输出 1 * n 维特征向量。

(3)反向传输调整权重。

实验的反向传输过程是卷积神经网络最复杂的地方,主要通过输出层、采样层和卷积层的最小化残差来调整权重和偏置,输出层的残差是输出值与类别值的误差,而中间各层的残差来源于下一层残差的加权和,实验最终通过 3 次迭代调整整个网络权重。

3.3　多通道语义特征融合

将 3.2 节中的三个通道提取出的 1 * n 维特征向量进行拼接,形成 3 * n 维多通道语义特征向量,然后按照 3.2 节的操作过程进行特征提取,最终形成 1 * n 维向量,在反向传输调整权重时的损失函数为式(5)。最后根据 SVM 分类器对多通道融合的语义特征进行分类,预测的准确率最高的事件类别即为对应的视频语义。

3.4　结果分析

表 2 和表 3 列出了文中提出的模型和其他对比模型在同一个 OA 数据集中每一个种类的识别准确率和平均准确率。

根据表 2,在 OA_1 数据集中,文中提出的 DVSM 模型在 10 种事件类别都取得了最高准确率,平均准确率为 69.4%。如表 3 所示,在 OA_2 数据集中,DVSM 模型 10 种事件类别中有 8 个准确率达到最高,DVSM 模型的平均准确率为 54.5%。对实验结果进行分析发现,识别错误的原因是实验的特征语义缺少事物特征、音频特征等,上述语义线索在人物语义识别也应被考虑和利用。

表 2　OA_1 视频场景数据集实验结果比较　%

Events	Xiaand Aggarwal (2013)	Ji et al. (2013)	DVSM
A answer the phone in office$_1$	12.5	40.0	80.0
A arrange in office$_1$	59.7	53.3	60.3
A eat in office$_1$	40.3	41.7	66.7
A go in to takesth and go out	48.6	51.7	60.0
A go to work in office$_1$	34.7	41.7	68.3
A look for in office$_1$	65.3	36.7	71.7
A wiping in office$_1$	63.9	66.7	76.7
A sleep in office$_1$	25.0	45	81.7
A take water in office$_1$	58.3	40.0	61.7
A wander in office$_1$	56.9	50.0	66.7
Accuracy	46.5	46.7	69.4

表 3　OA_2 视频场景数据集实验结果比较　%

Events	Xiaand Aggarwal (2013)	Ji et al. (2013)	DVSM
B ask C and away in office$_2$	12.5	39.6	62.3
B called C away in office$_2$	45.8	44.8	53.5
B and C carrysth in office$_2$	66.7	56.8	48.3
B chat with C in office$_2$	37.5	17.2	57.9
C deliversth to B in office$_2$	20.1	34.5	48.3
B eat and chat with C in office$_2$	50.0	35.8	46.6
Having-guest in office$_2$	37.5	34.1	55.2
C Seek-help B in office$_2$	16.7	44.8	56.1
Shaking-hands in office$_2$	41.7	32.8	51.7
B showingsth to C in office$_2$	37.5	29.3	64.6
Accuracy	36.6	37.0	54.5

4　结束语

利用基于视频场景的人物语义学习模型来完成视频中人物语义的识别。该模型使用卷积神经网络提取和融合人物身份、人物行为、上下文环境等通道语义信息,引入损失函数发现不同通道语义之间的潜在关联关系和精调整个网络学习参数,并通过 SVM 分类器完成识别人物语义任务。与现有的视频人物语义识别模型相比,提出的模型在特定数据集上识别的准确率较高,能够有效识别视频中人物的基本语义。

在人物语义识别中,视频中的音频、时序与一些逻辑知识信息都是识别视频中人物语义的重要线索[14-15],后续工作将研究如何在该模型中融合更多语

义线索,以提高语义识别的准确性。

参考文献:

[1] LI Yahui, CAI Cheng. Video segment retrieval based on affine hulls[C]//Proceeding of 2015 10th Asian control conference. [s. l.]:[s. n.],2015:1-6.

[2] 王 煜,周立柱,邢春晓. 视频语义模型及评价准则[J]. 计算机学报,2007,30(3):337-351.

[3] 吴 飞,刘亚楠,庄越挺. 基于张量表示的直推式多模态视频语义概念检测[J]. 软件学报,2008,19(11):2853-2868.

[4] PANG L, ZHU S, NGO C W. Deep multimodal learning for affective analysis and retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(11): 2008-2020.

[5] 沈 晴,班晓娟,常 征,等. 基于视频的人机交互中动作在线发现与时域分割[J]. 计算机学报,2015,38(12):2477-2487.

[6] KIM H, KIM J, OH T, et al. Blind sharpness prediction for ultra-high-definition video based on human visual resolution [J]. IEEE Transactions on Circuits & Systems for Video Technology, 2017, 27(5): 951-964.

[7] ZHU H, LIU Y, FAN J, et al. Video-based outdoor human reconstruction[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2017, 27(4): 760-770.

[8] 李志欣,施智平,李志清,等. 融合语义主题的图像自动标注[J]. 软件学报,2011,22(4):801-812.

[9] ATAN O, ANDREOPOULOS Y, TEKIN C, et al. Bandit framework for systematic learning in wireless video-based face recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(1): 180-194.

[10] KUMAR S H, SIVAPRAKASH P. New approach for action recognition using motion based features[C]//Proceedings of 2013 IEEE conference on information & communication technologies. Washington DC, USA: IEEE Computer Society, 2013: 1247-1252.

[11] LIANG Z, WANG X, HUANG R, et al. An expressive deep model for human action parsing from a single image[C]//Proceedings of 2014 IEEE international conference on multimedia and expo. Washington DC, USA: IEEE Computer Society, 2014: 1-6.

[12] ZHAN Y, DAI S, MA O Q, et al. A video semantic analysis method based on kernel discriminative sparse representation and weighted KNN[J]. The Computer Journal, 2015, 58(6): 1360-1372.

[13] ZHANG X, ZHANG H, ZHANG Y, et al. Deep fusion of multiple semantic cues for complex event recognition[J]. IEEE Transactions on Image Processing, 2016, 25(3): 1033-1046.

[14] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Washington DC, USA: IEEE Computer Society, 2015: 2625-2634.

[15] VENUGOPALAN S, HENDRICKS L A, MOONEY R, et al. Improving lstm-based video description with linguistic knowledge mined from text[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. [s. l.]: Association for Computational Linguistics, 2016: 1961-1966.

(上接第 52 页)

[12] KOREN Y, BORENSTEIN J. Potential field methods and their inherent limitations for mobile robot navigation[C]//Proceedings of IEEE conference on robotics and automation. Sacramento, CA, USA: IEEE, 1991: 1398-1404.

[13] 柳长安,鄢小虎,刘春阳,等. 基于改进蚁群算法的移动机器人动态路径规划方法[J]. 电子学报,2011,39(5):1220-1224.

[14] 杜鹏桢,唐振民,陆建峰,等. 不确定环境下基于改进萤火虫算法的地面自主车辆全局路径规划方法[J]. 电子学报,2014,42(3):616-624.

[15] 刘 玲,王耀南,况 菲,等. 基于神经网络和遗传算法的移动机器人路径规划[J]. 计算机应用研究,2007,24(2):264-265.

[16] 周 庆,牟 超,杨 丹. 教育数据挖掘研究进展综述[J]. 软件学报,2015,26(11):3026-3042.

[17] 云庆夏. 进化算法[M]. 北京:冶金工业出版社,2000.

[18] DOWNEY C, ZHANG Mengjie. Parallel linear genetic programming[C]//Proceedings of the 14th European conference on genetic programming. Torino, Italy: [s. n.], 2011: 178-189.

[19] KUO C S, HONG T P, CHEN C L. A knowledge-acquisition strategy based on genetic programming[J]. Cybernetics and Systems, 2008, 39(7): 670-683.

[20] ETEMADI H, ROSTAMY A A A, DEHKORDI H F. A genetic programming model for bankruptcy prediction: empirical evidence from Iran[J]. Expert Systems with Applications, 2009, 36(2): 3199-3207.

[21] ALAVI A H, GANDOMI A H. A robust data mining approach for formulation of geotechnical engineering systems [J]. Engineering Computations, 2011, 28(3): 242-274.

[22] OUANNES N, DJEDI N E, DUTHEN Y, et al. Gait evolution for humanoid robot in a physically simulated environment [M]//Intelligent computer graphics. [s. l.]: [s. n.], 2012: 157-173.

[23] GELLY S, TEYTAUD O, BREDECHE N, et al. A statistical learning theory approach of bloat[C]//Genetic and evolutionary computation conference. Washington DC, USA: IEEE, 2005: 1783-1784.