基于复合特征及深度学习的人群行为识别算法

袁亚军 李菲菲 陈 虬
(上海理工大学光电信息与计算机工程学院 上海 200093)

摘 要 分析人群行为的目的是更好地分析与管理人群运动的状态与趋势。针对人群行为的两种特征信息,提出了一种基于深度学习的人群行为识别方法。先将人群作为主要对象,通过前景提取方法来提取人群静态信息,利用人群运动的变化获取人群动态信息,借助卷积神经网络(CNN)模型学习这两种不同的人群行为特征,再综合这两种特征来分析常见的人群行为。同时,人群数据提取位置与间隔是影响人群行为分析的重要因素。实验结果表明,这两种人群特征能更好地描述空间维度上的人群状态和时间维度上的人群变化,合理的数据位置与数据间隔可以有效地提高人群信息的表达能力。最后将提出的方法与其他人群行为分析方法进行比较,定量与定性的实验结果验证了所提方法的有效性,同时也表明了所提方法能得到更优的混淆矩阵和更高的准确度。

关键词 人群行为识别,静态特征,动态特征,CNN,数据提取
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.046

Crowd Behavior Recognition Algorithm Based on Combined Features and Deep Learning

YUAN Ya-jun LEE Fei-fei CHEN Qiu
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract The target of analyzing crowd behavior is to better analyze and manage the state and tendency of crowd movement. This paper proposed a novel deep learning based crowd behavior recognition method by using two types of crowd behavior features. Firstly, the crowd is regarded as the main object, a foreground extraction method is used to extract the static information of crowd, and the dynamic information of crowd is obtained by the change of the crowd movement. Then two different crowd behavior characteristics are learned by using convolution neural network (CNN) model, so as to analyze crowd behaviors in the end. Additionally, the extraction location and interval of crowd data are crucial factors in the crowd behavior recognition. Experimental results show that two crowd characteristics can better describe crowd states on the spatial dimension and crowd changes on the temporal dimension. The rational data location and data interval can effectively improve the expression ability of crowd information. At last, this method was compared with other crowd behavior recognition algorithms. The quantitative and qualitative experimental results demonstrate the validity of the proposed method. Besides, better confusion matrix and higher precision can be obtained by this method.

Keywords Crowd behavior recognition, Static characteristic, Dynamic characteristic, CNN, Data extraction

1 引言

近年来,随着人们生活方式的改变,人口流动大、人群聚集频繁等人群活动现象越来越普遍,给社会带来了更多的安全隐患。人群聚集的场景[1]可能会引起人群拥挤甚至是人群踩踏现象,引发不必要的安全事故。因此人群行为分析[2]受到了研究者与社会的关注,特别是在城市安全与公共场所管理方面,合理地应用人群分析能够产生积极的影响。

人群[3]实际上是个体与群体的结合,它的主要信息来自于静态信息[4]与动态信息。人群的静态信息反映在人群视频的空间模块[5]上,而人群的运动信息反映在人群视频的时间模块上。相比于常见的图像分析,人群的动态过程为人群行为分析与理解提供了重要的信息来源,因为对于许多人群行为(人群聚集、分离等)需要在时间维度上分析其运动变化。如何分析场景中整个人群的行为研究关键,整个人群的状态决定着人群行为的变化。例如有组织的人群行走与无组织的人群行为,尽管大体上相似,都是人群行走的行为,但两者的区别在于有秩序和无秩序,在现实中产生的社会影响就是对前者可以进行平稳有序地管理,而后者可能产生人群拥堵等公共安全事件。这就是将人群行为识别视为更有价值的社会应用的基础的原因。

目前,该研究方向主要有两类不同的人群行为分析方法:

到稿日期:2018-05-12 返修日期:2018-09-13 本文受上海市高校特聘教授(东方学者)岗位计划(ES2015XX)资助。
袁亚军(1993-),男,硕士生,主要研究方向为计算机视觉与模式识别;李菲菲(1970-),女,博士,教授,主要研究方向为多媒体信息处理、图像处理与模式识别、信息检索,E-mail:feifeilee1701@163.com(通信作者);陈 虬(1972-),男,博士,教授,主要研究方向为图像处理与模式识别、计算机视觉、信息检索。

一类是利用传统的方法来提取特征并通过分类器来分类人群行为;另一类是利用深度学习的方法来分析人群行为。Shao 等[6]提出利用 4 种人群属性(集体性、一致性、稳定性与对抗性)作为人群行为的特征。在视觉角度上,这些人群属性是普遍存在于各种人群场景中的,其通过量化这 4 种人群描述符并利用支持向量机(SVM)来进行人群行为分类。Zhang 等[7]利用中级过滤器和 SVM 来提取全局特征,并结合方向梯度直方图(HOG)、霍夫变换与光流方向梯度直方图(HOF)等局部特征来理解人群。随着深度学习的发展,卷积神经网络(CNN)已经普遍应用于图像与视频的分析应用中。图像检索与分类、视频分析与分类等都会利用 CNN 来提取特征。近年来,随着深度学习的需要与硬件的发展,CNN 网络结构存在各种网络拓扑,几类常见的 CNN 网络结构被应用于人群分析领域。Ji 等[8]利用 3D CNN 模型来提取时间维度与空间维度上的人群特征,3D CNN 模型将 2D 输入扩展为 3D 输入,利用 3D 卷积来获取多个相邻视频帧的运动信息。Feichtenhofer 等[9]利用 two-stream CNN 模型将输入视频分为两个部分(时间部分与空间部分),时间部分以视频帧的各个像素点的移动来展示人群主体的运动,空间部分以单个视频帧的表面信息来反映视频中场景与人群的特征。该模型由两个相同的深度网络融合而成,这两个深度网络分别提取人群的时间信息和空间信息,从而学习人群特征。Shao 等[10-11]利用 S-CNN 模型将 3D 特征图分解为 2D 空间切片表示和 2D 时间切片表示,共产生了 3 个不同的维度分支。这些分支通过 3 个不同的 2D 时间滤波器和空间滤波器(xy-维度、xt-维度与 yt-维度)来学习不同维度方向上的人群外表特征与动态特征。

通过上述的分析可知,人群静态特征与动态特征在人群行为分析中是两类重要的人群特征。人群静态特征主要是以人群空间分布的形式呈现,反映了某一时间上人群的状态;而人群动态特征主要是以人群时间变化的形式表现,反映了某段时间内人群的运动变化。本文将这两类人群特征与 CNN 模型相结合来进行人群行为识别研究。

2 人群特征提取

视频[12]是人群行为表达的载体,视频中人群特征的提取是人群行为识别与分类的基础,其流程如图 1 所示。

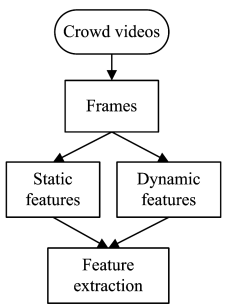


图 1 特征提取流程图

Fig. 1 Flowchart of feature extraction

从人群视频中抽取图像帧,通过图像获取人群的静态信息和动态信息,利用特征提取器来提取人群行为特征。

2.1 数据提取

人群视频通常包含几十张甚至几百张图像帧。在人群行为分析中,不需要提取视频中的所有图像帧,因为这不仅会增加处理与分析的复杂性,还会消耗大量的时间,浪费硬件内存。文献[13-14]只从每个人群视频中提取前 30 帧图像。由于部分人群视频在初始阶段存在不稳定的模糊数据的干扰,而人群视频中间阶段的帧图像干扰较小,能够得到更加稳定、有效的帧数据,因此本文采用在人群视频中提取中间位置的帧图像进行处理的方式。在动态信息提取过程中,帧间隔的大小会影响动态特征的表达效果。若帧的间隔太小,则动态特征的表达效果就会下降;反之,会引起动态信息数据的冗余,从而影响动态特征的学习。

2.2 人群静态特征

人群视频的静态特征[15-16]包含人群空间分布、人群密度、背景等。在人群行为分析中,背景对研究的影响相对较小,因为人群行为分析的主要对象是人群本身;相反,背景比较少的人群信息增加了特征的维度,从而加大了算法的复杂度。因此,本文方法以通过检测的人群图像来分析静态特征的方式代替当前大多数文献通过视频帧的图像来学习人群静态特征的方式,检测的人群图像最大化地保留了人群的信息。

前景提取方法用于获取人群图像,前景提取的方法有很多类,每一类方法都有不同的优缺点,因此其应用场合也不同。常见的前景提取方法有帧差法、高斯混合模型与背景相减法。在实际应用中,这些方法都有限制条件。帧差法是通过相邻视频帧的灰度差来提取前景目标,而当人群是静止不动、缓慢运动时或者视频的分辨率较低时,相邻视频帧的灰度变化较小,很难通过差值来提取出前景目标。高斯混合模型采用多个单高斯分布来表示视频帧的各个像素点的变化,很难应用于高密度人群的场景。尽管应用背景相减法建立背景的模型并实时更新背景是较困难的,但本文数据库中的每个人群视频只有一个场景,且场景中只存在一个稳定不变的背景图像,因此利用背景相减法来提取人群图像是合理的。

本文背景相减法的流程如图 2 所示。

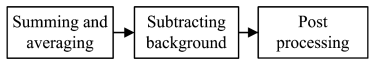


图 2 背景相减法的流程

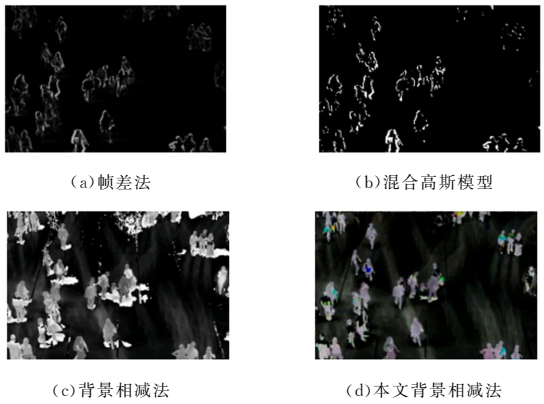
Fig. 2 Flowchart of background subtraction method

背景相减法的流程主要分为 3 个部分:1)将所有视频帧的像素对应求和,再平均化得到每个视频的背景图像;2)将提取的帧图像减去对应的背景图像,得到对应视频帧的人群图像;3)后期处理,包括阈值处理与滤波处理。为了在人群行为分析中尽可能地保留颜色信息,整个背景相减的过程都是在 RGB 三通道下处理的,这也不同于常见的背景相减的灰度处理过程。由于未处理的人群图像很粗糙,且存在类似于椒盐噪声的干扰,因此利用中值滤波可以有效地消除该类噪声。阈值处理的原理是:

$$p=\begin{cases} p, & p<t \\ 0, & p\geqslant t \end{cases} \tag{1}$$

其中,p 是人群图像中每个像素点的值,参数 t 是阈值。

前景提取方法的效果如图 3 所示。



(a)帧差法　(b)混合高斯模型　(c)背景相减法　(d)本文背景相减法

图 3　前景提取方法的效果对比

Fig. 3　Comparison of results obtained by foreground extraction methods

从图 3 中可以看出，帧差法(见图 3(a))与混合高斯模型(见图 3(b))只能提取出运动人群的轮廓，而背景相减法(见图 3(c)和图 3(d))可以提取出运动的人群，尽可能地保留人群信息，本文在此基础上进行了平滑处理，还保留了人群的颜色信息，提升了图像的人群静态特征的学习能力。

2.3　人群动态特征

人群视频[17]的动态特征包括人群运动方向、人群运动速度等。光流法是人群理解中常见的动态特征表达方法，它利用一段时间内图像序列中像素点的变化与相邻视频帧的关联性来获取前一帧与当前帧的关系，最后计算出相邻帧的运动物体的信息。LK 光流法是广泛使用的光流估计的差分方法，根据光流法的前提假设(亮度恒定、微小运动)得到基本的光流约束方程：

$$\frac{\partial I}{\partial x}V_x+\frac{\partial I}{\partial y}V_y+\frac{\partial I}{\partial t}=0 \tag{2}$$

其中，$\frac{\partial I}{\partial x},\frac{\partial I}{\partial y},\frac{\partial I}{\partial t}$表示图像在$(x,y,t)$点的梯度，$V_x,V_y$ 表示该点的光流值。根据空间一致性的前提假设，图像每个像素点的光流约束方程组成一个超定方程，联合求解出最优解，即图像的光流值。文献[18]均利用这种方式来表达人群的动态信息，但是得到的实验效果并不是很好，因此本文采用将其他动态信息叠加到光流图上的方式来提高人群动态特征的表达能力。

人群运动图是在一段时间内通过运动像素点的位置变化来表达人群运动的方向与速度。其通过部分运动检测点在不同图像上的位置变化来表达人群运动的方向，并以矢量线标注。动态信息图如图 4 所示。

(a)人群运动图　(b)光流叠加图

图 4　动态信息图

Fig. 4　Dynamic information images

2.4　特征提取器

相比于传统的特征提取方法，CNN[19-21]的自学习能力和参数共享机制能够更好地处理高维数据，特征的分类效果也较好。CNN 的网络结构在总体上是类似于双流网络的，但其输入数据、融合位置与方法不同。网络的输入数据是人群图像与叠加光流图，融合方式是在全连接层前通过 Concat 层来连接这两个通道。相比于其他深度网络模型(VGG，GoogleNet，ResNet)，AlexNet 模型的参数较少，计算复杂度较小，占用的内存更小，硬件要求也较低，不易过拟合。CNN 网络结构的 2 个通道都采用 AlexNet 网络结构，AlexNet 模型包含 5 个卷积层与 3 个全连接层。具体的 CNN 网络结构如图 5 所示。

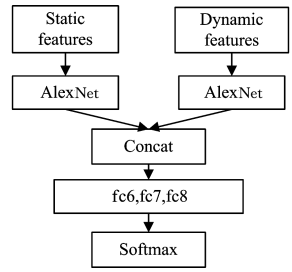


图 5　CNN 网络结构

Fig. 5　Structure of CNN network

3　实验结果与分析

本文在两类人群分析(人群状态分析与人群行为分类)的应用上，通过定量与定性实验来证明所提方法的有效性。

3.1　人群数据库

实验数据库选用 CUHK 人群视频库，其含有 474 个人群视频和 215 个不同的场景。该数据是在不同的环境下获取的，如街道、购物商场、机场与公园，并且人群的密度和拍摄视角也各不相同。尽管数据库中的视频有不同的长度，但人群行为分析只需要从每个人群视频中提取部分帧图像作为实验数据来验证方法的有效性。数据库中视频的类别标签已明确标注，不需要人工标注。

3.2　评价指标

利用混淆矩阵和准确度来评价图像分类结果。混淆矩阵是由真实类别和预测类别组成的 4 个部分计算而得，这 4 个部分是真正类(TP)、假负类(FN)、假正类(FP)与真负类(TN)。混淆矩阵如表 1 所列。

表 1　混淆矩阵

Table 1　Confusion matrix

Predict \ Actual	Positive class	Negative class
Positive class	True Positive(TP)	False Positive(FP)
Negative class	False Negative(FN)	True Negative(TN)

准确度(Accuracy)是对分类模型的总体判断，是总量中正确预测的比例，计算方法如下：

$$Accuracy=\frac{TP+TN}{TP+FP+FN+TN} \tag{3}$$

3.3　人群状态分析

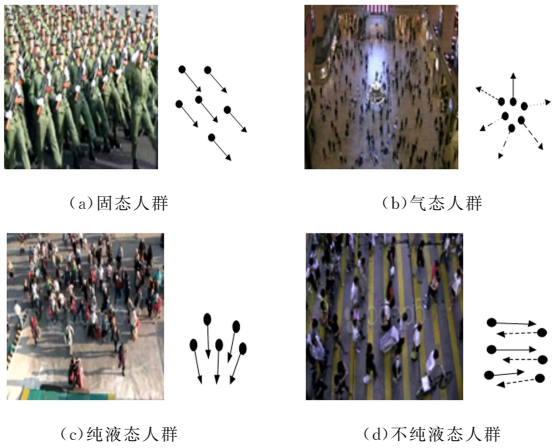
人群状态由多个物理因素决定，包括人群密度与人群成员之间的相互关系，不同的状态对应着潜在的物体模型。下面给出 4 种人群状态的描述。

1)固态人群:人群集体向相同的方向运动,他们之间的相互位置不变,内部作用力始终约束着人群的运动。

2)气态人群:人群向不同的方向运动,没有形成与其他人一致的行为。

3)纯液态人群:人群向相同的方向移动,但是由于缺乏内部作用力的影响,他们的相对位置不断改变。

4)不纯液态人群:类似于纯液态人群,但其存在与其他人群的碰撞。



(a)固态人群 (b)气态人群

(c)纯液态人群 (d)不纯液态人群

图 6 4 种人群状态

Fig. 6 Four crowd states

根据人群数据库中存在的标签将 928 张人群帧图像标记为真实的人群状态,人群帧图像的具体分布是 290 张固态人群图、130 张气态人群图、348 张纯液态人群图、160 张不纯液态人群图。根据本文方法分别提取人群图像与叠加光流图,并将它们作为 CNN 模型的输入数据。其中一半数据用于训练,其余用于测试,数据设置与文献[6]一致。

实验结果的混淆矩阵如表 2 所列,其中,气态人群和纯液态人群的分类结果普遍好于固态人群和不纯液态人群。将人群状态的分类结果与文献[6]中利用人群属性产生的分类结果(见表 3)进行比较可知,本文方法的分类准确度都有所提高,实验结果优于文献[6]的实验结果。

表 2 本文分类人群状态的混淆矩阵

Table 2 Confusion matrix for classifying crowd states in this paper

Predict \ Actual	Gas	Solid	Pure fluid	Impure fluid
Gas	0.95	0	0	0.05
Solid	0	0.729	0.243	0.028
Pure fluid	0.06	0	0.994	0
Impure fluid	0.012	0	0.425	0.563

表 3 文献[6]分类人群状态的混淆矩阵

Table 3 Confusion matrix for classifying crowd states in reference [6]

Predict \ Actual	Gas	Solid	Pure fluid	Impure fluid
Gas	0.75	0.06	0.16	0.03
Solid	0.02	0.66	0.29	0.03
Pure fluid	0.09	0.24	0.52	0.15
Impure fluid	0.14	0.08	0.32	0.46

利用静态特征的单网络结构与本文双网络结构进行分类的实验结果如表 4 所列。双网络结构的准确度为 84.6%,比单网络结构的准确度高出 2%,也优于文献[1]中 60%的准确度。

表 4 不同网络结构的结果比较

Table 4 Comparison of different network structures

CNN network	*Accuracy*/%
static feature CNN	82.6
static and dynamic feature CNN	84.6

3.4 人群行为分类

根据标签将数据库中 474 个人群视频划分为 8 类,具体如表 5 所列。

表 5 人群行为类别

Table 5 Categories of crowd behavior

	Crowd behavior categories
1	Highly mixed pedestrian walking
2	Crowd walking following a mainstream and well organized
3	Crowd walking following a mainstream but poorly organized
4	Crowd merge
5	Crowd split
6	Crowd crossing in opposite directions
7	Intervened escalator traffic
8	Smooth escalator traffic

这 8 种类别在人群视频中是很常见的,部分类别有利于公共场所中人群的管理与交通管控。例如,人群合并与人群交叉可能会引起交通拥堵、人群踩踏等公共安全事件;人群行为分析也对商场出入口区域扶梯的平稳运行起到一定的作用,可以避免阻塞、冲突与潜在的安全隐患。

类别 1 是人群以不同的行为向不同的方向行走,类别 2 和类别 3 是人群大体上沿着主流方向行走。类别 2 是有秩序地行走,他们的相对位置稳定,很难发生应急突发事件;而类别 3 是无组织地运动,他们的相对位置可以随时变化,很有可能发生人群拥挤现象。类别 4—类别 8 通常出现在人群管理与公共交通场所,类别 4 是人群合并,有可能会发生人群拥挤现象;类别 7 是需要干预的扶梯交通,这说明该场景已经出现人群阻塞现象。

3.4.1 不同数据提取方式的比较

在人群视频中,提取相同数量、不同数据位置的人群静态特征图,将这些数据作为 AlexNet 深度网络模型的输入。提取视频中间数据产生的实验结果和提取视频前端数据产生的实验结果如表 6 所列。提取中间数据受到的分类准确度为 82.6%,而提取前端数据的分类准确度为 78%,这说明视频中间数据受到的干扰较小,对人群特征的表达能力更强。

表 6 数据位置的实验结果比较

Table 6 Comparison of experimental results of data location

Data location	*Accuracy*/%
front	78.0
median	82.6

同时,不同的帧数量与帧间隔也影响着人群实验的结果。分别提取不同数量与不同间隔的图像数据作为 AlexNet 深度网络模型的输入,图 7 和图 8 分别展示了不同的帧数量与不同的帧间隔得到的人群分类结果。

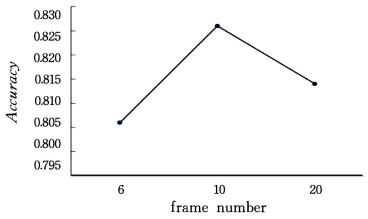


图 7　帧数的实验结果比较

Fig. 7　Comparison of experimental results of frame number

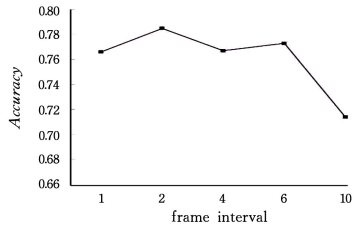


图 8　帧间隔的实验结果比较

Fig. 8　Comparison of experimental results of frame interval

在帧数的实验中，10 帧的实验效果为 82.6%，而 6 帧的实验效果为 80.6%，帧数过少使得数据量不足，会导致特征学习能力下降；而帧数过多使得特征冗余，消极的特征会影响人群行为分类结果，如 20 帧的准确度反而降低为 81.4%。

在帧间隔的实验中，帧间隔为 1，2，4，6 的实验效果相差较小，帧间隔为 2 的效果最佳，准确度达到了 78.5%。

3.4.2　人群行为分类结果

以上实验结果表明，选择合适的数据参数（中间位置、10 帧图像且帧间隔为 2）可以得到更好的人群分类结果。本文从数据库的每个人群视频中提取对应条件的人群静态数据与动态数据，共产生 4740 张人群静态特征图与相对应的动态特征图，将这些特征图作为 CNN 双网络结构的输入来进行特征学习。其中，80%的数据用于训练，其余 20%的数据用于测试。本文方法的实验结果与其他方法的实验结果的比较如表 7 所列，可知本文方法的准确度为 80.5%，优于其他对比方法。

表 7　本文算法与其他算法的实验结果比较

Table 7　Comparison of experimental results of proposed algorithm with other algorithms

Algorithms	*Accuracy*/%
Shao et al[1]	70.0
Burney et al[11]	70.0
Ours	80.5

3.5　算法分析

人群行为分析的主体是人群本身，人群信息是人群行为的表达。原始数据中的背景信息对人群行为分析的积极影响较小，反而增加了信息维度。因此提取人群信息可以提高人群行为特征学习的效率。表 8 列出了人群信息对算法的影响。

表 8　人群信息对算法的影响

Table 8　Effect of crowd information on this algorithm

CNN input	Accuracy/%
original data	77.4
crowd data	80.5

表 8 中的数据表明，与原始数据相比，人群信息的提取提升了人群行为分类的效果。

由于传统的背景减法易受到环境光照的影响，因此利用伽马校正（gamma correction）来分析光照对本文算法的影响。当 *gamma* 值小于 1 时，图像的整体亮度得到了提升。表 9 列出了光照对算法的影响。

表 9　光照对算法的影响

Table 9　Effect of light on this algorithm

gamma	*Accuracy*/%
0.2	81.9
0.4	79.2
0.6	78.3
0.8	79.6
—	80.5

实验数据显示，不同的光照（*gamma* 值）对算法的影响在±2%之内，对本文算法的影响较小。

由于光流法计算量较大，本文利用 10 帧数据来代表视频数据，减少了信息数据的计算量，算法的实时性如表 10 所列。

表 10　本文算法的运行时间

Table 10　Running time of proposed algorithm

Validation	Period	Time/s
95videos	prior disposal	609.1
(950frames)	prediction	3435.7

结束语　本文提出了基于深度学习的人群行为识别方法，利用前景提取方法凸显人群的静态信息，用光流法表达人群的动态信息，并结合运动图来增强人群的动态表达能力。在此基础上，使用 CNN 双网络模型提取人群的静态特征与动态特征，联合分析人群的状态与行为。同时，本文分析了数据提取方式对人群行为识别的影响。实验表明，本文方法优于以人群属性为特征的方法，在 CUHK 数据库上得到了更优的混淆矩阵和更高的准确度。在未来的研究工作中，将考虑两种特征之间的依存关系，通过两者的联系来分析人群行为。

参 考 文 献

[1] SHAO J，KANG K，CHEN C L，et al. Deeply learned attributes for crowded scene understanding[C]// International Conference on Computer Vision and Pattern Recognition. 2015：4657-4666.

[2] PALANISAMY G，MANIKANDAN T T. Group Behaviour Profiling for Detection of Anomaly in Crowd[C]// International Conference on Technical Advancements in Computers and Communications. 2017：11-15.

[3] RODRIGUES F，LOURENCO M，RIBEIRO B，et al. Learning Supervised Topic Models for Classification and Regression from Crowds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，PP(99)：1-1.

[4] YI S，LI H，WANG X. Pedestrian Behavior Modeling from Stationary Crowds with Applications to Intelligent Surveillance[J]. IEEE Transactions on Image Processing，2016，25(9)：4354-4368.

[5] SENGUPTA S，WANG H，BLACKBURN W，et al. Spatial information in classification of activity videos[C]// 2015 Federa-

ted Conference on Computer Science and Information Systems. 2015:145-153.

[6] SHAO J,CHEN C L,WANG X. Learning Scene-Independent Group Descriptors for Crowd Understanding[J]. IEEE Transactions on Circuits & Systems for Video Technology,2017,27(6):1290-1303.

[7] ZHANG C,KANG K,LI H,et al. Data-Driven Crowd Understanding:A Baseline for a Large-Scale Crowd Dataset[J]. IEEE Transactions on Multimedia,2016,18(6):1048-1061.

[8] JI S,YANG M,YU K. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2013,35(1):221-231.

[9] SIMONYAN K,ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in Neural Information Processing Systems,2014,1(4):568-576.

[10] SHAO J,CHEN C L,KANG K,et al. Slicing Convolutional Neural Network for Crowd Video Understanding[C]// International Conference Conference on Computer Vision and Pattern Recognition. 2016:5620-5628.

[11] JING S,CHEN C L,KAI K,et al. Crowded Scene Understanding by Deeply Learned Volumetric Slices[J]. IEEE Transactions on Circuits & Systems for Video Technology,2017,27(3):613-623.

[12] KARPATHY A,TODERICI G,SHETTY S,et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// International Conference Computer Vision and Pattern Recognition. 2014:1725-1732.

[13] SHAO J,LOY C C,WANG X. Scene-Independent Group Profiling in Crowd[C]// International Conference on Computer Vision and Pattern Recognition. 2014:2227-2234.

[14] BURNEY A,SYED T Q. Crowd Video Classification Using Convolutional Neural Networks[C]// International Conference on Frontiers of Information Technology. 2017:1255-1259.

[15] YI S,WANG X. Profiling stationary crowd groups[C]// International Conference on Multimedia and Expo. 2014:1-6.

[16] YI S,LI H,WANG X. Understanding pedestrian behaviors from stationary crowd groups[C]// International Conference on Computer Vision and Pattern Recognition. 2015:3488-3496.

[17] SENST T,EISELEIN V,SIKORA T. A local feature based on lagrangian measures for violent video classification[C]// International Conference on Imaging for Crime Prevention and Detection. 2015:1-6.

[18] FEICHTENHOFER C,PINZ A,ZISSERMAN A. Convolutional Two-Stream Network Fusion for Video Action Recognition[C]// International Conference on Computer Vision and Pattern Recognition. 2016:1933-1941.

[19] GOMEZ-DONOSO F,GARCIA-GARCIA A,GARCIA-RODRIGUEZ J,et al. LonchaNet:A sliced-based CNN architecture for real-time 3D object recognition[C]// International Joint Conference on Neural Networks. 2017:412-418.

[20] MARSDEN M,MCGUINNESS K,LITTLE S,et al. Fully Convolutional Crowd Counting On Highly Congested Scenes[J]. arXiv preprint arXiv:1612.00220.

[21] SHI Y,TIAN Y,WANG Y,et al. Sequential Deep Trajectory Descriptor for Action Recognition with Three-Stream CNN[J]. IEEE Transactions on Multimedia,2017,19(7):1510-1520.