JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
12
VOL.25

ISSN1006-8961
CN11-3758/TB

数字浮雕建模 *P2647*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第12期（总第296期）

2020年12月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

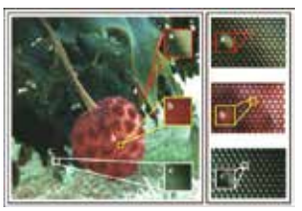
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

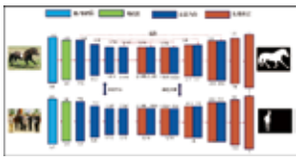
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年12月 第25卷 第12期（总第296期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

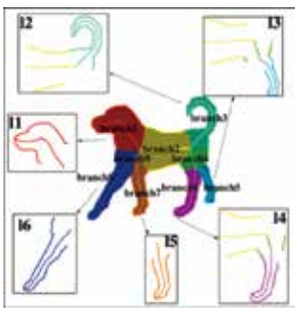
目次



光场显著性检测研究综述(第2465页)



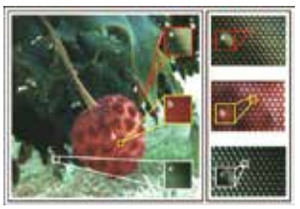
融合注意力机制与知识蒸馏的孪生网络压缩(第2563页)



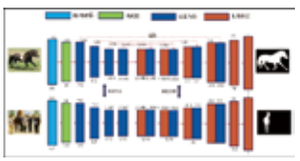
面向可解释性的物体拓扑结构骨架表征方法(第2587页)

Volume 25, Number 12
Published December 16, 2020

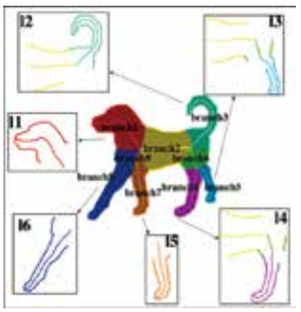
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Review of saliency detection on light fields(P2465)



Combining attention mechanism and knowledge distillation for Siamese network compression (P2563)



Skeleton characterization of object topology toward explainability (P2587)

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2020)12-2517-13

论文引用格式: Ma M, Li Y B, Wu X Q, Gao J F and Pan H P. 2020. Human action recognition in videos utilizing key semantic region extraction and concatenation. Journal of Image and Graphics,25(12):2517-2529(马淼,李贻斌,武宪青,高金凤,潘海鹏. 2020. 关键语义区域链提取的视频人体行为识别. 中国图象图形学报,25(12):2517-2529)[DOI:10.11834/jig.200049]

关键语义区域链提取的视频人体行为识别

马淼[1],李贻斌[2],武宪青[1],高金凤[1],潘海鹏[1]

1. 浙江理工大学机械与自动控制学院, 杭州　310018; 2. 山东大学控制科学与工程学院, 济南　250100

摘　要: **目的**　视频中的人体行为识别技术对智能安防、人机协作和助老助残等领域的智能化起着积极的促进作用,具有广泛的应用前景。但是,现有的识别方法在人体行为时空特征的有效利用方面仍存在问题,识别准确率仍有待提高。为此,本文提出一种在空间域使用深度学习网络提取人体行为关键语义信息并在时间域串联分析从而准确识别视频中人体行为的方法。**方法**　根据视频图像内容,剔除人体行为重复及冗余信息,提取最能表达人体行为变化的关键帧。设计并构造深度学习网络,对图像语义信息进行分析,提取表达重要语义信息的图像关键语义区域,有效描述人体行为的空间信息。使用孪生神经网络计算视频帧间关键语义区域的相关性,将语义信息相似的区域串联为关键语义区域链,将关键语义区域链的深度学习特征计算并融合为表达视频中人体行为的特征,训练分类器实现人体行为识别。**结果**　使用具有挑战性的人体行为识别数据集 UCF(University of Central Florida)50 对本文方法进行验证,得到的人体行为识别准确率为 94.3%,与现有方法相比有显著提高。有效性验证实验表明,本文提出的视频中关键语义区域计算和帧间关键语义区域相关性计算方法能够有效提高人体行为识别的准确率。**结论**　实验结果表明,本文提出的人体行为识别方法能够有效利用视频中人体行为的时空信息,显著提高人体行为识别准确率。

关键词:人机交互;深度学习网络;人体行为关键语义信息;人体行为识别;视频关键帧

Human action recognition in videos utilizing key semantic region extraction and concatenation

Ma Miao[1], Li Yibin[2], Wu Xianqing[1], Gao Jinfeng[1], Pan Haipeng[1]

1. Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou　310018, China;
2. School of Control Science and Engineering, Shandong University, Jinan　250100, China

Abstract: **Objective** Human action recognition in videos aims to identify action categories by analyzing human action-related information and utilizing spatial and temporal cues. Research on human action recognition are crucial in the development of intelligent security, pedestrian monitoring, and clinical nursing; hence, this topic has become increasingly popular among researchers. The key point of improving the accuracy of human action recognition lies on how to construct distinctive features to describe human action categories effectively. Existing human action recognition methods fall into three categories: extracting visual features using deep learning networks, manually constructing image visual descriptors, and combining manual construction with deep learning networks. The methods that use deep learning networks normally operate convolution and pooling on small neighbor regions, thereby ignoring the connection among regions. By contrast, manual construction

收稿日期:2020-02-18;**修回日期**:2020-03-13;**预印本日期**:2020-03-20

基金项目:浙江省自然科学基金项目(LQ19F030014,LQ18F030011);浙江理工大学青年创新专项(2019Q035)

Supported by: Natural Science Foundation of Zhejiang Province, China (LQ19F030014, LQ18F030011)

methods often have strong pertinence and poor adaptability to specific human actions, and its application scenarios are limited. Therefore, some researchers combine the idea of handmade features with deep learning computation. However, the existing methods still have problems in the effective utilization of the spatial and temporal information of human action, and the accuracy of human action recognition still needs to be improved. Considering the above problems, we research on how to design and construct distinguishable human action features and propose a new human action recognition method in which the key semantic information in the spatial domain of human action is extracted using a deep learning network and then connected and analyzed in the time domain. **Method** Human action videos usually record more than 24 frames per second; however, human poses do not change at this speed. In the computation of human action characteristics in videos, changes between consecutive video frames are usually minimal, and most human action information contained in the video is similar or repeated. To avoid redundant computations, we calculate the key frames of videos in accordance with the amplitude variation of the image content of interframes. Frames with repetitive content or slight changes are eliminated to avoid redundant calculation in the subsequent semantic information analysis and extraction. The calculated key frames contain evident changes of human body and human-related background and thus reveal sufficient human action information in videos for recognition. Then, to analyze and describe the spatial information of human action effectively, we design and construct a deep learning network to analyze the semantic information of images and extract the key semantic regions that can express important semantic information. The constructed network is denoted as Net1, which is trained by transfer learning and can use continuous convolutional layers to mine the semantic information of images. The output data of Net1 provides image regions, which contain various kinds of foreground semantic information and region scores, which represent the probability of containing foreground information. In addition, a nonmaximal suppression algorithm is used to eliminate areas that have too much overlap. Afterward, the key semantic regions are classified into person and nonperson regions, and then the position and proportion of person regions are used to distinguish the main person and the secondary persons. Moreover, object regions that have no relationship with the main person are eliminated, and only foreground regions that reveal human action-related semantic information are reserved. Afterward, a Siamese network is constructed to calculate the correlation of key semantic regions among frames and concatenate key semantic regions in the temporal domain. The proposed Siamese network is denoted as Net2, which has two inputs and one output; Net2 can be used to mine deeply and measure the similarity between two input image regions, and the output values are used to express the similarity. The constructed Net2 can concatenate the key semantic regions into a semantic region chain to ensure the time consistency of semantic information, and express human action change information in time domain more effectively. Moreover, we tailor the feature map of Net1 using the interpolation and scaling method, in order to obtain feature submaps of uniform size. That is, each semantic region chain corresponds to a feature matrix chain. Given that the length of each feature matrix chain is different, the maximum fusion method is used to fuse the feature matrix chain and obtain a single fused matrix, which reveals one kind of video semantic information. We stack the fused matrix from all feature matrix chains together and then design and train a classifier, which consists of two fully connected layers and a support vector machine. The output of the classifier is the final human action recognition result for videos. **Result** The UCF (University of Central Florida) 50 dataset, a publicly available challenging human action recognition dataset, is used to verify the performance of our proposed human action recognition method. In this dataset, the average human action recognition accuracy of the proposed method is 94.3%, which is higher than that of state-of-the-are methods, such as that based on optical flow motion expression (76.9%), that based on a two-stream convolutional neural network (88.0%), and that based on SURF (speeded up robust features) descriptors and Fisher encoding (91.7%). In addition, the proposed crucial algorithms of the semantic region chain computation and the key semantic region correlation calculation are verified through a control experiment. Results reveal that the two crucial algorithms effectively improve the accuracy of human action recognition. **Conclusion** The proposed human action recognition method, which uses semantic region extraction and concatenation, can effectively improve the accuracy of human action recognition in videos.

Key words: human-machine interaction; deep learning network; key semantic information of human action; human action recognition; video key frame

0 引 言

随着人工智能技术的飞速发展，人机交互变得愈发频繁，各种智能体及智能机器人进入了人们的日常生产与生活，为人类活动提供了极大的便利。人机交互程度的不断加深使得如何让机器更好地识别人体行为成为一项研究热点（Schydlo 等，2018；Huang 和 Mutlu，2016）。人体行为识别的研究在行为监控与辅助、患者监护与康复，视频内容分析与检索等领域有重要意义，是人工智能及人机交互中的一项重要技术（Mishra 等，2019；Mocanu 等，2018）。

根据视频中人体行为特征提取方法及计算侧重点，人体行为识别方法分为 3 类：1）利用深度学习网络获得人体行为表达特征进行人体行为识别；2）利用图像视觉信息设计并构造人体行为表达特征，进行人体行为识别；3）将深度学习网络与手工构造特征相结合进行人体行为识别。

深度学习网络以单帧图像为输入，能够深度挖掘图像特征（Simonyan 和 Zisserman，2014a；Yang 等，2015），常用于实现人体行为识别。Zhu 等人（2018）提出一种利用两个并行深度学习网络对视频的彩色图像信息与运动信息分别处理的方法。两个深度学习网络结构不同，分别独立地深度挖掘视频的空间信息与运动信息，然后将挖掘到的特征表达向量相连接，得到既能描述图像空间信息又能描述时间变化信息的人体行为识别描述子，实现人体行为的有效分类与识别。但是，该方法是以整幅图像作为输入，图像中既含有人体行为相关的前景区域，又包含无关的背景区域，仅适用于视频中人体运动幅度在图像中占比足够大的情况。Xiao 等人（2019）考虑到与整幅图像尺度相比，人体行为幅度过小时容易被忽略，提出了一种利用 3 层空间金字塔对人体行为图形进行不同尺度表达，然后用卷积神经网络计算不同尺度对应行为特征的方法，该方法能解决大图像直接送入深度学习网络造成的细微行为特征被忽视的问题。

然而，有学者认为，上述直接使用深度学习网络对视频帧中的信息进行处理的方法仅对小范围邻域进行卷积和池化，忽视了图形中重要的空间相关联信息（Gkioxari 和 Malik，2015；Peng 等，2014），因而提出了基于图像信息有目的地构造人体行为特征的计算方法，能够更有效地识别不同的人体行为。Kardaris 等人（2016）将视频时间域计算得到的密度轨迹（Wang 和 Schmid，2013）与视频图像空间域的视觉词汇序列相结合，并统计归纳为人体行为特征描述子，用来对人体行为类别特征进行表达。该方法在密度轨迹中提取了人体行为的时间一致性信息，在视觉词汇序列中汇聚了人体行为的空间信息，从而实现视频中人体行为的有效分类。Mishra 等人（2019）利用图像空间信息计算图像中聚类点之间的距离与几何关系，根据聚类点在视频中的变化进行统计，并将空间几何信息与有向图相结合，从而计算视频帧中人体行为的表达特征，对视频中的人体行为进行识别。该方法构造的聚类集群特征能够对帧内图像特征有效表示，同时具有帧间运动特征的表达能力，从而实现对视频中人体行为类别的有效识别。

然而，上述基于图像信息的方法依赖手工构造的特征，虽然在处理人体空间信息时具有目的性更加明确的优点，但该类方法对特定人体行为的针对性较强，适应性较差，应用场景受到限制。

因此，有学者提出，将依赖手工制造的特征目的明确的优点与深度学习方法擅于归纳的特点相结合（罗会兰和张云，2019），能够更有针对性地使用深度学习方法，从而得到更有表征能力的人体行为表达特征。Zheng 等人（2018）提出首先使用深度学习网络获取视频中可能的人体位置区域，然后在该区域使用手工构造的方法有针对性地计算图像中人体的轮廓及角点特征，并对得到的图像视觉特征进行统计处理，从而获得更为明确的人体形态轮廓信息，然后对该信息进行特征融合并训练分类器，实现视频中人体行为的识别。Ma 等人（2018）考虑到人体行为类别特征通常体现在人体与背景物体的交互上，而仅对人体行为前景区域进行处理的方法会忽略人体前景与图像背景进行交互的信息，因此提出一种分别处理人体前景区域和全图区域的人体行为识别方法，该方法首先使用基于手工构造图像特征的方法得到人体前景位置区域，然后将前景区域和全图区域分别用深度学习框架进行分析，从而获得包含人物交互信息的人体行为识别表达特征。

通过上述分析可知，将手工构造特征的针对性与深度学习网络的适应性相结合，是有效提高人体行为识别准确率的一种新思路。

为进一步有效提高视频中人体行为识别的准确率,本文设计了一种基于空间维提取关键语义区域并在时间维进行相关性计算的人体行为识别方法,具体贡献如下:

1)提出一种从图像中分析并提取关键语义区域的方法。设计深度学习网络对视频中的关键语义信息进行分析和提取,将完整图像根据不同语义信息分解为多个图像区域。每个图像区域表达一种语义信息,从而能够更细致更有针对性地对人体行为空间信息进行表达,有效提高了人体行为识别的准确率。

2)提出一种使用孪生神经网络将帧间关键语义区域串联的方法。构造孪生神经网络对帧间关键语义区域进行相关性计算,并将帧间语义信息一致的区域进行串联,得到帧间关键语义区域链,从而保证语义信息的一致性,更有针对性地对人体行为时间域变化信息进行表达,有效提高了人体行为识别的准确率。

3)使用迁移学习的方法训练深度学习网络,使算法具有较强的适应性,极大减少了针对某一具体行为的训练过程计算量,克服了现有方法针对具体行为需要进行额外训练的问题。

1 视频图像中关键语义区域的计算

为了更好地对图像中的人体行为空间信息进行解析和利用,需要对能够表达图像内容的关键语义区域进行提取。关键语义区域能够更有针对性地表达人体行为类别特征。

1.1 计算视频关键帧

视频是由图像帧序列构成的,由于人眼的视觉暂留效应,多幅图像以超过24帧/s的速度播放(Zhi和Cooperstock,2012)便会成为连续的视频。然而,在针对视频中人体行为特征进行计算的任务中,往往连续视频帧间变化不大,包含的人体行为信息类似或重复。为避免重复计算,在挖掘视频中的人体行为语义信息及行为表达特征前,对视频帧进行筛选,剔除内容变化不明显的图像,提取能够反映人体行为变化的视频关键帧。

图像内容由各个像素进行表达,视频图像通常为3通道彩色图像,为方便计算,首先将彩色图像灰度化,得到尺寸为 $a \times b$ 的单通道图像。然后,将该图像分割为 $n_0 \times n_0$ 的网格,当图像尺寸不能整除网格数 n_0 时,考虑到图像关键信息通常更靠近图像正中,因此在两端均匀空出不能整除的像素数。得到的图像网格如图1所示。按照图1的网格对图像像素进行统计,每个网格记为 $\boldsymbol{c}_{i,j}$,统计 $\boldsymbol{c}_{i,j}$ 中的像素值之和,记为 $S_{i,j}$,具体为

$$S_{i,j} = \sum_{x=(i-1)\cdot a/n_0+1}^{i\cdot a/n_0} \sum_{y=(j-1)\cdot b/n_0+1}^{j\cdot b/n_0} (g(x,y)) \tag{1}$$

式中,$g(x,y)$ 表示位于图像帧中 (x,y) 位置处的像素灰度值;a/n_0 和 b/n_0 分别为每个网格内 x 方向和 y 方向的网格数目。$n_0 \times n_0$ 个 $S_{i,j}$ 构成矩阵 $\boldsymbol{S}$。

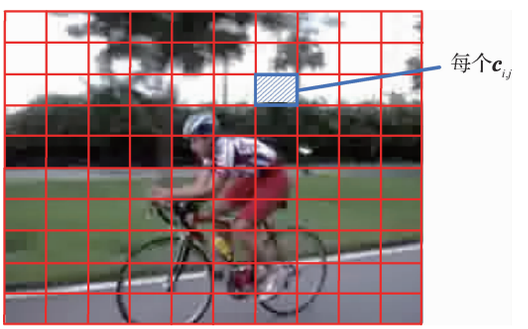


图1 对像素变化进行分区域统计

Fig. 1 Count the variance of pixels by region

为获得图像内容变化较为明显的图像关键帧,首先将第1帧作为关键帧(即 $p=1$),然后逐帧计算第 q 帧图像与关键帧 p 之间的网格差异,具体为

$$\boldsymbol{D}^{p,q} = \boldsymbol{S}^p - \boldsymbol{S}^q \tag{2}$$

式中,$\boldsymbol{S}$ 是由式(1)中的 $S_{i,j}$ 构成的 $n_0 \times n_0$ 的矩阵,上标 p 和 q 表示图像帧的索引号。计算得到的 $\boldsymbol{D}^{p,q}$ 是一个 $n_0 \times n_0$ 的矩阵,矩阵中的每一个元素代表对应网格位置的灰度值变化量。灰度图像的像素值取值为 $g(x,y) \in [0,255]$ 之间的整数,网格内像素值之和 $S_{i,j}$ 变化的阈值 δ_c 为

$$\delta_c = 4 \cdot \frac{a}{n_0} \cdot \frac{b}{n_0} \tag{3}$$

表示像素值在[0,255]范围内平均变化4个灰度值即视为图像区域内容有明显变化。

阈值 δ_c 为像素值之和 $S_{i,j}$ 的变化阈值,用 δ_c 对 $\boldsymbol{D}^{p,q}$ 矩阵表示的网格差异程度进行约束,若 $\boldsymbol{D}^{p,q}$ 矩阵中存在大于阈值 δ_c 的区域数目 $N_D > 10$,则认为视频帧 q 是相对于视频帧 p 的关键帧,如图2所示。遍历整个视频帧序列,可得到视频中能够表达图像内容明显变化的关键帧序列,从而减小后续图像处理的冗余计算量。

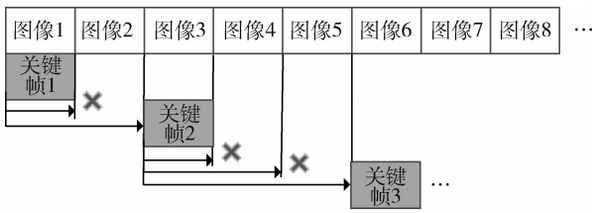


图 2　通过逐帧对比计算获得图像关键帧

Fig. 2　Obtain key frames through comparing frame by frame

1.2　根据图像语义信息提取关键区域

在得到视频中的关键帧后，分析每幅图像表达的语义信息和最能表征图像关键内容的语义信息，分割出蕴含该信息的图像区域。

本文设计并使用一种卷积神经网络，实现对图像中语义信息的挖掘。该网络是通过使用单幅静态图像和图像中含有的特定物体的标签及位置进行端到端训练得到的（Ren 等，2017），网络中卷积层的功能是挖掘图像内容中蕴含的信息，并对感兴趣的区域及其语义信息进行分析。图 3 展示了本文提出的用于提取图像关键语义区域的网络，为方便叙述，将该网络结构记为 Net1。

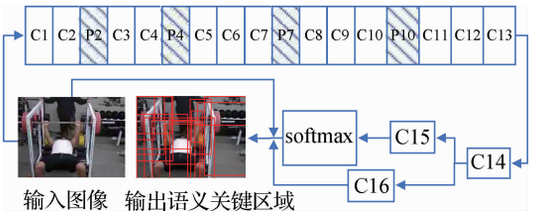


图 3　用于提取图像关键语义区域的网络 Net1

Fig. 3　Framework of Net1 for extracting key semantics areas

在 Net1 网络中，C 表示卷积层，P 表示池化层，softmax 表示使用的逻辑回归模型。卷积层 C1—C13 的作用是深度挖掘图像中含有的语义信息（Simonyan 和 Zisserman，2014b），使用的卷积核为 3 × 3，填充为 1，卷积步长为 1；池化层使用的池化核为 2 × 2，填充为 0，卷积步长为 2。每次卷积操作保持图像大小不变，每次池化使特征图像尺寸变为上一层的 1/2，因此 C13 的输出特征图尺寸为输入图像的 1/16。

C14 的卷积核仍为 3 × 3，填充为 1，卷积步长为 1，目的是进一步集中特征图的语义信息。C15 和 C16 分别使用 9 × 2 和 9 × 4 个 1 × 1 的核（Ren 等，2017）对 C14输出的特征图进行全卷积，目的是使用 9 种不同形状的目标框对特征图进行遍历计算，从而得到相应的语义信息位置及其得分。网络对生成的候选区域进行相应的裁剪过滤后（Girshick，2015），通过 softmax 判断并保留图像帧中的关键语义区域。

Net1 网络的参数是在 VOC2007（Visual object classes challenge 2007）数据集上通过迁移学习得到的（Everingham 等，2007），为了更加有效地使用 VOC2007 数据集的注释信息，Net1 网络需连接一个分类器，使用数据集中注释的物体位置及标签信息进行端到端的学习（Girshick 等，2014），因此，Net1 中网络参数的学习过程为弱监督学习。Net1 网络的输出为多个候选关键语义区域 $\{\boldsymbol{r}_i\}$ 及其对应的包含前景信息的概率 $\{s_i\}$。概率 $\{s_i\}$ 将用于关键语义区域 $\{\boldsymbol{r}_i\}$ 的进一步计算及处理。

在使用 Net1 提取关键语义区域时，首先对视频关键帧图像进行缩放，将长和宽中较短的边长缩放至 600 像素，以适应 Net1 网络中的卷积及池化操作。Net1 网络的输出为多个候选关键语义区域，将所得区域按得分排序，并按顺序计算交并比，剔除相似的区域，使用的限制条件为

$$\begin{cases} \text{剔除 } \boldsymbol{r}_a \text{ 区域} & \dfrac{\boldsymbol{r}_a \cap \boldsymbol{r}_b}{\boldsymbol{r}_a \cup \boldsymbol{r}_b} > 0.6 \\ \text{保留 } \boldsymbol{r}_b \text{ 区域} & \text{其他} \end{cases} \tag{4}$$

式中，$\boldsymbol{r}_a$ 区域为已保留的关键语义区域，$\boldsymbol{r}_b$ 为待评价的关键语义区域。使用非极大值抑制算法后，将视频中每帧图像中含有的候选关键语义区域数量 N_r 控制在上限 $N_\delta = 200$ 以内。这些关键语义区域的集合定义为 $\{\boldsymbol{r}_k\}$，其中，k 表示关键区域索引号。

图 4 展示了使用 Net1 网络提取到的关键语义区域样例，其中每个区域代表一种语义信息。

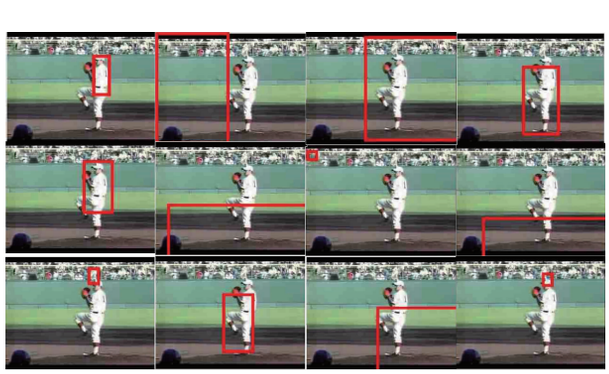


图 4　Net1 网络提取的关键语义区域

Fig. 4　Key semantic regions extracted by Net1

2 视频中关键语义区域的串联

2.1 视频中关键语义区域串联的方法

在得到关键帧图像中的候选关键语义区域后，考虑到图像区域之间含有重复信息，进一步对由同一帧图中提取的 N_r 个关键区域进行筛选，以减少冗余计算。

为进行人体行为识别，需要从图像的人体行为关键语义区域提取人体行为表达特征，而人体行为主要表现在内容充实的人体区域和与人体相关的区域。较小的区域通常含有的内容较少，通过分析，剔除 $\{\boldsymbol{r}_k\}$ 中较小的候选语义区域，即

$$\begin{cases}\text{剔除 } \boldsymbol{r}_i \text{ 区域} & \max(\text{size}(\boldsymbol{r}_i)) < 20 \text{ 像素} \\ \text{保留 } \boldsymbol{r}_i \text{ 区域} & \text{其他}\end{cases} \tag{5}$$

人体所在的图像区域包含了视频中人体行为识别的重要信息，因此，进一步使用 VOC2007 数据集（Everingham 等，2007）中的人体类别，训练人体区域分类器支持向量机（support vector machine，SVM）（Girshick 等，2014），可计算出候选关键语义区域中的人体区域及其概率。

考虑到一幅图像中存在多人的情况，将图像中的不同人体按人体区域分类器的输出概率排序，依次计算人体区域间的重叠情况，剔除重叠区域大于比例阈值 δ_r 的区域，仅保留满足阈值 δ_r 条件的区域，具体为

$$\frac{\boldsymbol{r}_i^p \cap \boldsymbol{r}_j^p}{\min_s(\boldsymbol{r}_i^p, \boldsymbol{r}_j^p)} < \delta_r \tag{6}$$

式中，$\boldsymbol{r}_i^p$ 与 $\boldsymbol{r}_j^p$ 表示同一帧图像中任意两个人体区域，$\min_s(\cdot)$ 表示求两个区域中面积较小者，δ_r 为比例阈值，实际计算时比例阈值取 $\delta_r = 0.2$。

通过上述操作，当视频中含有多个人体时，可分析并计算出每帧图像中相互独立的人体区域，如图 5 所示。

含有多个人体的图像中通常只有一个主要行为人，当视频含有多人时，需要区分主人物和副人物。为此，使用深度学习网络框架 Net2 对帧间两区域进行相关性计算，将帧间的同一人物区域串联得到串联区域链。综合人物串联结果，取占面积更大且出现位置更靠近图像中心位置的人物作为主人物，该人体区域链记为 $\{\boldsymbol{r}^\alpha\}$；其余人物作为副人物，每个副人物对应的串联区域链记为 $\{\boldsymbol{r}_i^\beta\}$。

图 5 用分类器筛选出人体区域

Fig. 5 Select areas of the person utilizing classifier

人体行为的关键信息为人体及人与物交互的语义区域。因此，在每帧图像中，针对主人物的区域位置，剔除与主人物区域交集为 $\boldsymbol{\phi}$ 的语义区域。剔除后，将每帧图像中的人与物交互的关键语义区域分别按时间顺序使用深度学习网络框架 Net2 进行相关性计算，串联成关键语义区域链，记为 $\{\boldsymbol{r}_i^\gamma\}$，其中 i 为不同语义区域链的索引号。

深度学习网络框架 Net2 是一种孪生神经网络，如图 6 所示，能够深度挖掘两图像区域的相似性并进行度量。

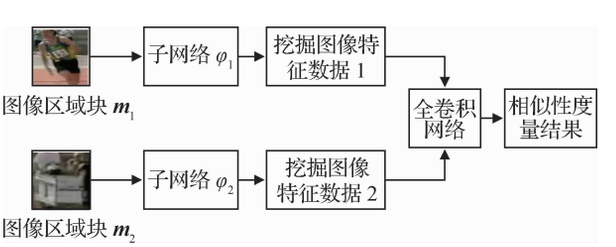


图 6 用于关联帧间信息的网络 Net2

Fig. 6 Framework of Net2 for associating information among frames

2.2 用于关联帧间语义信息的网络架构

为反映视频中的人体行为语义信息的变化，需要对视频帧间的关键语义区域进行相关性分析。

现有的视频中人体行为识别方法通常通过计算光流图来获得视频帧间人体行为的连续性信息（Tu 等，2018；Chéron 等，2015），然而光流图仅能表达帧间图像的相对运动，无法对图像内容及语义信息的关联性进行分析。针对该问题，本文构造并使用一种孪生神经网络对视频中的关键语义区域进行相关性

计算，将视频帧间表达相同语义信息的图像区域串联，从而对时间域中的人体行为语义信息进行分析。

孪生神经网络有两个输入一个输出，可用来计算两幅图像内容的相似性（He 等，2018；Krizhevsky 等，2012），具体为

$$f(\boldsymbol{m}_1,\boldsymbol{m}_2) = g(\varphi_1(\boldsymbol{m}_1),\varphi_2(\boldsymbol{m}_2)) \tag{7}$$

式中，$\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 表示两幅输入图像；$f(\cdot)$ 表示用孪生神经网络求取其相似性，该操作可等价为用 $\varphi_1(\cdot)$ 和 $\varphi_2(\cdot)$ 两个深度学习网络分别挖掘 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 两幅图像的特征，然后利用 $g(\cdot)$ 进行相似性判别。

在实际操作中，根据具体问题，$\varphi_1(\cdot)$ 和 $\varphi_2(\cdot)$ 可选择相同或不同网络（Bertinetto 等，2016）。本文使用孪生神经网络计算两幅结构相同图像的关键语义区域，因此 $\varphi_1(\cdot) = \varphi_2(\cdot) = \varphi(\cdot)$ 为相同网络，即式(7)可进一步表示为

$$f(\boldsymbol{m}_1,\boldsymbol{m}_2) = a[\varphi(\boldsymbol{m}_1) * \varphi(\boldsymbol{m}_2)] + b \tag{8}$$

式中，$*$ 操作代表全卷积，a 和 b 分别表示线性变换的比例系数和偏置。这样，该网络可输出评价 $\boldsymbol{m}_1$ 与 $\boldsymbol{m}_2$ 之间相似程度的分值。

在使用 Net2 计算两区域间的相似性时，需将输入图像块 $\boldsymbol{m}_1$ 与 $\boldsymbol{m}_2$ 调整为相同尺寸 $n_\varphi \times n_\varphi$。图 6 中的子网络 φ 的结构如图 7 所示。

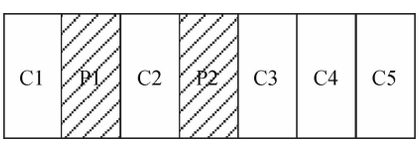


图 7　子网络 φ 的结构

Fig. 7　Framework of the sub-net φ

图 7 中，子网络 φ 由 5 个卷积层和 2 个池化层构成，其中，卷积层 C1 使用的卷积核为 11×11，步长为 2，C2 的卷积核为 5×5，步长为 1；池化层 P1 和 P2 使用的核为 3×3，步长为 2；卷积层 C3—C5 使用的卷积核均为 3×3，步长为 1。当子网络 φ 的输入数据结构为 $127 \times 127 \times 3$ 时，输出的深度挖掘特征数据结构为 $6 \times 6 \times 128$。

Net2 网络的作用是计算两幅图像区域的相似程度，为增加该网络的鲁棒性，网络参数是在不相关的数据集 ILSVRC（large scale visual recognition challenge）数据集（Russakovsky 等，2015）上通过迁移学习得到的。

2.3　帧间关键语义区域串联的效果

在使用孪生神经网络将图像关键语义区域串联时，从第 1 帧中的各个关键区域开始向后计算，相似性度量结果最大的两个关键语义区域视为表达语义信息相同的区域，可将其串联；若与后一帧中所有区域的相似性度量结果最大值仍小于阈值 δ_b，则视为无可串联区域，停止串联。即

$$\begin{cases} \text{无可串联区域} & \max(f(\boldsymbol{m}_i^l,\boldsymbol{m}_j^{l+1})) < \delta_b \\ \text{可串联} & \text{其他} \end{cases} \tag{9}$$

当某一帧 t 中出现无法与前一帧相串联的区域时，则作为新的关键语义区域向后计算。遍历整个视频所有语义信息关键区域进行相似性计算，考虑到视频中已剔除图像内容变化不大的帧，因此若某语义区域出现在连续 3 个关键帧内，则视其为有效语义信息，即保留长度大于 3 的关键区域串联结果，这样便得到多条语义区域链 $\{\boldsymbol{r}_i\}$。

在 1.2 节中提到，图像帧中提取的每个语义信息关键区域均对应一个由 Net1 网络计算得到的属于该区域前景的概率 $\{s_i\}$。计算每条语义区域链的概率均值，可得

$$p_i = \frac{1}{n_i}\sum_l s_i^l \tag{10}$$

式中，n_i 表示该语义区域链的长度，l 表示语义区域链中每个元素的索引号。

对所有语义区域链按对应的概率均值排序。根据 2.1 节的分析，语义区域链包括主人物语义区域链 $\{\boldsymbol{r}^\alpha\}$、副人物语义区域链 $\{\boldsymbol{r}_i^\beta\}$ 和其他语义区域链 $\{\boldsymbol{r}_i^\gamma\}$。因此，根据该优先级依次保留主人物语义区域链 1 条、全部副人物语义区域链 n_β 条和较优的其他语义区域链 $(N-1-n_\beta)$ 条，即从一段人体行为视频中提取出 N 条语义区域链，实际操作时，取 $N = 16$。图 8 展示了使用本文提出的关键区域串联方法计算得到的部分语义区域链样例。

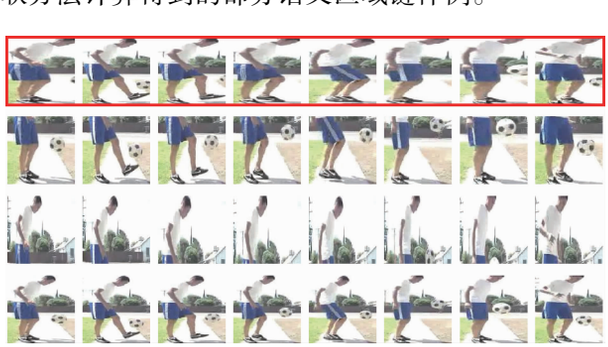


图 8　由关键区域串联得到的语义区域链样例

Fig. 8　Examples of the semantic region chains obtained by concatenating key regions

3 人体行为特征的构造及识别

得到人体行为语义信息的关键区域链后，需要设计合适的特征提取方法，从关键区域链中提取出能够有效表达人体行为类别的信息，从而实现人体行为的有效识别，具体过程如图 9 所示。

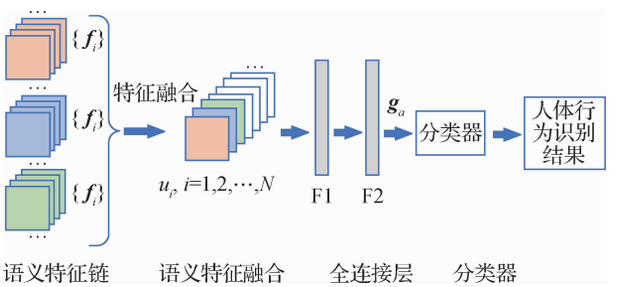


图 9 构造人体行为识别特征

Fig. 9 Construct human action recognition features

通过关键语义区域的串联计算，从一段视频中可提取到 N 条语义区域链 $\{\boldsymbol{r}_i\}$。将每个关键语义区域对应到 Net1 网络的 C14 层特征图的相应部分进行切割，并将切割出的特征子图进行插值和缩放，使特征子图尺寸统一为 $n_f \times n_f$，得到的语义特征链记为 $\{\boldsymbol{f}_i\}$，其中，$i = 1,\cdots,N$。

由于各语义特征链 $\{\boldsymbol{f}_i\}$ 的长度不同，需将每条 $\{\boldsymbol{f}_i\}$ 中含有的信息融合为同尺寸数据。为此，使用最大值融合法（Wang 等，2016b）将每条 $\{\boldsymbol{f}_i\}$ 中的语义信息及其变化进行融合，即求取每条 $\{\boldsymbol{f}_i\}$ 中 $n_f \times n_f$ 的每个位置的最大值，具体为

$$\boldsymbol{\mu}_i = \max_{x \in n_f \times n_f} (\{\boldsymbol{f}_i\}|_x) \tag{11}$$

式中，$\boldsymbol{\mu}_i$ 为新构成的特征，反映了第 i 种语义信息在整个视频中的变化。然后，使用两层全连接层 F1 和 F2 将融合的语义特征变为长度为 4 096 的向量。全连接层的优势是在提取并强化特征向量的同时对空间位置的变化具有鲁棒性。一个全连接层能够表达特征的线性映射关系，但会丢失部分非线性关系，而连续使用两个全连接层能够提取线性及非线性特征信息，从而得到更加有效的人体行为表达特征。

将第 2 个全连接层 F2 的输出作为视频中人体行为的表达特征，记为 $\boldsymbol{g}_a$，该特征含有从深度学习网络挖掘到的视频中与人体相关的语义信息，因此能够更好地表达视频中的人体行为。然后，使用支持向量机（SVM）作为分类器（Chang 和 Lin，2011）对人体行为的表达特征 $\boldsymbol{g}_a$ 进行分类，从而实现视频中人体行为的有效识别。

4 实验结果与分析

4.1 UCF50 数据集

为验证本文提出的人体行为识别方法的有效性，使用 UCF（University of Central Florida）50 数据集（Reddy 和 Shah，2013）进行验证。

UCF50 数据集是美国中佛罗里达大学机器视觉课题组研究人员从 YouTube 视频中采集得到的，含有 50 个行为类别，每个行为类别包括不同人物在不同环境的 100 余段行为视频，共 6 618 段。

UCF50 数据集中的人体行为类别多样，来源于真实环境，背景多变且复杂，是公认的具有挑战性的人体行为数据集（Peng 等，2016；Nazir 等，2018）。图 10 是数据集中几类行为的视频帧。

图 10 UCF50 数据集中的行为实例

Fig. 10 Instances of actions in the UCF50 dataset

4.2 实验结果与分析

在本文算法中，用于提取关键语义区域的 Net1 网络和用于实现关键区域串联的 Net2 网络分别是在 VOC2007 和 ILSVRC 数据集上通过迁移学习训练得到的。因此，Net1 和 Net2 网络不需要针对 UCF50 数据集进行训练，仅训练分类器即可，使得本文方法在训练过程的计算量大幅减少，从而对不同场景及不同人体行为具有较强的适应性。

在使用 UCF50 数据集对本文人体行为识别方法进行验证时，使用留一交叉验证法（Reddy 和 Shah，2013）对分类器进行训练，计算各类人体行为识别准确率的均值作为人体行为识别的结果。人体行为识别的准确率是某类行为中能够正确识别的视频数占该类行为视频总数的百分比。某类行为 A 的准确率 P_A 计算为

$$P_A = \frac{C(\beta_A \mid \alpha_A)}{C(\alpha_A)} \times 100\% \tag{12}$$

式中，条件 α_A 表示实际为 A 类行为；条件 β_A 表示由算法识别为 A 类行为；运算符 C 表示计算符合条件的视频数量。

使用本文人体行为识别方法对 UCF50 数据集中的 50 类人体行为进行识别，得到的每类行为的人体行为识别准确率如图 11 所示。

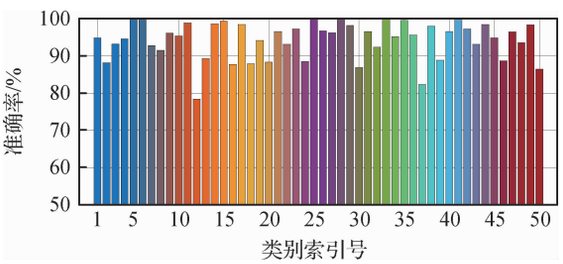


图 11 在 UCF50 数据集上的人体行为识别准确率

Fig. 11 Human action recognition results obtained on UCF50

从图 11 可以看出，本文人体行为识别算法对每类行为的识别能力不同。当视频图像中包含的大部分语义信息与人体行为类别的关联度较高时，使用本文人体行为识别方法能够实现准确识别。例如，拳击（第 33 类行为）的图像语义信息能够较好地表达挥动拳击手套的动作，因此该类行为的识别准确率较高，识别率为 100% 。而当图像中存在大量与人体行为类别关联性较小的语义信息时，冗余和无关的语义信息会对人体行为的识别结果产生干扰。例如，跳高（第 12 类行为）的图像语义信息包含环境复杂的田径场上各种干扰信息，因此行为识别准确率受到影响，仅为 78.4% 。

综合分析本文人体行为识别方法在 UCF50 数据集上的实验结果，可以看出，本文算法对人体行为识别的准确率普遍较高，均值为 94.3% ，其中准确率在 90% 以上的类别占总类别的 76% ，说明本文人体行为识别方法能够对视频中的人体行为进行有效识别。

4.3 对比实验与分析

为进一步说明本文人体行为识别方法的有效性，将本文方法与现有的前沿人体行为识别方法进行对比，使用 UCF50 数据集进行验证，结果如表 1 所示。

表 1 本文方法与对比方法在 UCF50 数据集上的人体行为识别准确率

Table 1 Proposed method compared with the state-of-the-art methods for human action recognition accuracy on UCF50 dataset

/%

方法	人体行为识别准确率
Wang 等人（2016a）	91.7
Nazir 等人（2018）	93.4
Reddy 和 Shah（2013）	76.9
Mocanu 等人（2018）	88.0
Hu 等人（2019）	90.8
本文	**94.3**

注：加粗字体为最优结果。

表 1 中，Wang 等人（2016a）使用 SURF（speeded up robust features）算子描述人体行为的关键角点，每帧中的 SURF 算子通过光流相关联，再用 Fisher 向量进行编码，融合视频各帧中的描述子，从而对人体行为进行表达。该方法使用的图像描述子能够独立表达检测到的图像特征点，却无法对各描述子之间的关联性进行分析和表达，需要有针对性地训练复杂的编码和特征融合。而本文算法能够对图像的语义信息进行时空域分析，从而更具针对性地对人体行为相关区域进行深度信息提取。本文方法与 Wang 等人（2016a）的方法相比，人体行为识别准确率提高了 2.6% 。

Nazira 等人（2018）使用 3D Harris 角点描述子和 3D SIFT（scale-invariant feature transform）特征提

取法计算训练集中每种行为的视频特征,然后通过聚类算法为每类行为训练并创建表达特征词包,实现视频中的人体行为类别识别。该方法对人体行为特征进行提取和描述依赖于图像梯度的变化,在背景复杂多变的现实场景中不能有效挖掘与人体行为相关的信息,但该方法为每种行为单独构造特征描述子和特征词包,在牺牲训练过程计算量的同时弥补了特征提取方法目的性不足的缺点,人体行为识别准确率为93.4%。本文算法提取图像中的语义信息并挖掘和分析与人体相关的语义特征,在特征计算方面更具有目的性,在大幅减少分类器训练过程计算量的前提下,人体行为识别准确率为94.3%,优于Nazira等人(2018)的方法。

Reddy和Shah(2013)通过计算光流图区分运动的前景和静止的背景,分别计算特征描述子,将前景信息与背景信息相融合用于人体行为识别。该方法虽然考虑了前景与背景中的语义信息,但用光流图粗略地区分前景与背景得到的分割并不准确,人体行为识别准确率为76.9%。与本文方法相比,该方法没有考虑前景与背景交互的语义信息,因此人体行为识别准确率明显低于本文方法。

Mocanu等人(2018)使用两个卷积神经网络分别对视频图像帧序列及相邻帧间光流图序列进行处理和计算,将两个网络的输出特征进行融合并训练分类器,以实现人体行为识别任务。该方法以完整的图像帧及光流图作为输入,采用深度学习网络提取特征,将得到的特征融合后直接用于训练分类器,因此对视频中的人体行为信息的分析及利用不充分,人体行为识别准确率为88%。与之相比,本文方法考虑了视频中不同区域的人体行为之间含有不同的语义信息,及视频中语义信息的时间连续性和每种语义信息随时间的变化,识别结果明显占优。

Hu等人(2019)得到的人体行为识别准确率为90.8%。该方法使用两个深度学习网络分别对视频的彩色图像和光流图中的人体行为特征进行提取,然后对得到的深度学习特征矩阵进行时空特征池化和时空金字塔池化,再对得到的数据进行主成分分析,将得到的特征向量用于训练分类器,实现人体行为识别。该方法对运动信息的提取依赖于光流图,然而光流图只能表达相邻帧之间的运动,不能表达有语义信息的空间域变化。而本文方法通过孪生神经网络将表征关键语义信息的区域在时间区域串联,可以表达连续语义信息的时间域变化,从而有针对性地计算人体行为特征,人体行为识别准确率优于Hu等人(2019)的识别结果。

通过上述对比分析可以看到,与现有的先进方法相比,本文方法在对人体行为语义信息分析与提取方面具有优越性,能够有针对性地计算对人体行为识别贡献最大的关键语义区域的图像特征,并通过使用孪生神经网络有效地分析出关键语义区域在时间域的关联性及随时间的变化,从而对人体行为表达信息进行更有针对性的分析和计算。因此,本文方法能够更好地实现视频中的人体行为识别。

4.4 算法有效性分析

本文人体行为识别方法中有两个关键算法,视频帧中关键语义区域计算和帧间关键语义区域相关性计算。

为进一步验证本文人体行为识别方法的有效性,针对本文方法中的两个关键算法分别进行有效性实验,结果如表2所示。

表2 关键算法的有效性实验

Table 2 Experiments on the effectiveness of key algorithms

/%

方法	人体行为识别准确率
不使用关键语义区域计算	56.2
不使用帧间区域相关性计算	75.8
本文	**94.3**

注:加粗字体为最优结果。

表2第1行是视频帧中关键语义区域计算的关键算法有效性验证对比实验结果。该方法不使用本文提出的关键语义区域计算方法,而是将完整的人体行为视频帧图像作为计算对象,使用Net1网络C14层的输出结构作为每帧的人体行为识别特征,然后使用与本文方法相同的特征融合及训练分类器的方法进行计算,人体行为识别准确率仅为56.2%,这是由于该方法对完整视频帧中的所有图像区域予以相同的关注度,即无法对人体所在区域有针对性地计算,也不能对人与物交互的语义信息进行剖析。对比实验结果说明,本文人体关键语义信息提取的关键算法能够增强对人体行为关键区域的有效利用,提高算法的针对性,从而有助于提高人体行为识别准确率。

表 2 第 2 行是帧间关键语义区域相关性计算的关键算法有效性验证对比实验结果。该对比实验在使用本文算法得到关键语义区域后，不使用孪生神经网络 Net2 进行串联，而是通过计算帧间光流图得到的运动信息和颜色直方图相似性度量将各区域串联。具体地，使用 Net1 计算第 1 帧中属于前景语义区域概率较大的 N 个关键语义区域，利用光流运动在时间域连接每帧中的颜色直方图最匹配的区域，从而得到与本文所提方法类似的语义区域链。然后用本文使用的特征融合及分类器训练的方法构成完整的人体行为识别策略，人体行为识别准确率为 75.8% 。与之相比，本文提出的使用帧间语义区域相关性计算的方法得到的人体行为识别准确率提高了 18.5% ，这是由于光流图仅能表达帧间的相对运动，颜色直方图仅能表达图像表观信息的颜色一致性，而本文提出的利用孪生神经网络进行相关性计算的方法能够表达图像区域语义信息，更好地关联视频序列中所含语义信息相似且连贯的图像区域，从而更有效地分析人体行为信息并提取人体行为表达特征，因此有效提高了人体行为识别准确率。

通过上述对本文关键算法的分析与验证，进一步证明了本文人体行为识别方法能够有效提高人体行为识别的准确率。

4.5 实验结果实例展示

图 12 是使用本文方法对 UCF50 数据集中的人体行为进行识别的结果实例。图 12(a) 是识别正确的 4 个样例。可以看出，在图像语义信息能够较好地对人体行为的关键动作进行描述时，本文人体行为识别方法效果较好。例如，拳击(图 12(a)) 的语义信息能够较好地表达挥动拳击手套的信息，室内攀岩(图 12(a)) 的语义信息能够较好地表达人体贴附于攀岩墙上的信息。图 12(b) 是识别错误的两个样例。跳高(图 12(b)) 受复杂背景干扰和拍摄模糊问题的影响，导致图像语义信息识别度不高，误识别为掷标枪。划船(图 12(b)) 由于拍摄方向问题使得图像语义信息产生歧义，无法获得船的形状及运动特征，误识别为水上摩托。

图 12 实验结果实例展示

Fig. 12 Demonstration of experimental results

((a) samples of right recognition; (b) samples of wrong recognition)

5 结 论

针对现有人体行为识别方法难以同时对视频中人体行为的时间域和空间域特征进行有效描述的问题，本文提出一种在空间域对人体行为关键语义区域进行提取，并在时间域进行相关性计算的识别方法，提高了视频中人体行为识别准确率。具体地，

1)根据帧间图像内容计算视频关键帧,避免后续语义信息分析与提取时产生冗余计算。2)构造深度学习网络 Net1 和 Net2。Net1 对视频帧中的语义信息进行分析,有针对性地解析和利用人体行为空间信息;Net2 对帧间关键语义区域进行相关性度量,分析语义信息时间域特征。3)计算关键语义区域链的深度学习特征并进行融合,构造出人体行为表达特征,实现对人体行为更细致更有针对性的表达,从而有效提高视频中人体行为识别的准确率。

本文通过使用有挑战性的人体行为识别公开数据集 UCF50 对所提方法进行了验证实验,得到的人体行为识别准确率为 94.3% 。通过对比实验,表明本文所提人体行为识别方法的检测准确率优于其他 5 种现有的前沿人体行为识别方法。本文方法包括两个关键算法,对其进行验证实验,进一步证明了本文方法的有效性。实验结果和分析表明,与其他方法相比,本文方法通过有针对性地对人体行为时空信息进行分析和表达,有效提高了视频中人体行为识别的准确率。此外,本文构造的深度学习网络 Net1 和 Net2 均使用迁移学习的方法进行训练,在提高网络适应性的同时,解决了现有人体行为识别方法针对具体行为需要大量额外训练的问题。但是本文人体行为识别方法需要提取视频中人体行为类别的关键语义信息,当视频中的语义信息杂乱、拍摄模糊或拍摄角度使图像内容产生歧义时,识别有效性会受到影响。

未来的研究工作将致力于将人体行为识别方法应用于智能家居系统,进一步提高人机交互的灵活性,实现人机自然交互。

参考文献(References)

Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional Siamese networks for object tracking//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherland: Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]

Chang C C and Lin C J. 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3): 1-27 [DOI: 10.1145/1961189.1961199]

Chéron G, Laptev I and Schmid C. 2015. P-CNN: pose-based CNN features for action recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3218-3226 [DOI: 10.1109/iccv.2015.368]

Everingham M, Van Gool L, Williams C K I, Winn J and Zisserman A. 2007. The PASCAL visual object classes challenge 2007 (VOC2007) results [EB/OL]. [2019-12-22]. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/cvpr.2014.81]

Gkioxari G and Malik J. 2015. Finding action tubes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 759-768 [DOI: 10.1109/cvpr.2015.7298676]

He A F, Luo C, Tian X M and Zeng W J. 2018. A twofold siamese network for real-time object tracking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE: 4834-4843 [DOI: 10.1109/cvpr.2018.00508]

Hu H, Liao Z and Xiao X. 2019. Action recognition using multiple pooling strategies of CNN features. Neural Processing Letters, 50(1): 379-396 [DOI:10.1007/ s11063-018-9932-3]

Huang C M and Mutlu B. 2016. Anticipatory robot control for efficient human-robot collaboration//Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interaction. Christchurch, New Zealand: IEEE: 83-90 [DOI: 10.1109/HRI.2016.7451737]

Kardaris N, Pitsikalis V, Mavroudi E and Maragos P. 2016. Introducing temporal order of dominant visual word sub-sequences for human action recognition//Proceedings of 2016 IEEE International Conference on Image Processing. Phoenix, USA: IEEE: 3061-3065 [DOI: 10.1109/ICIP.2016.7532922]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 1097-1105

Luo H L and Zhang Y. 2019. Semantic segmentation method with combined context features with CNN multi-layer features. Journal of Image and Graphics, 24(12): 2200-2209 (罗会兰, 张云. 2019. 结合上下文特征与 CNN 多层特征融合的语义分割. 中国图象图形学报, 24(12): 2200-2209) [DOI: 10.11834/jig.190087]

Ma M, Marturi N, Li Y B, Leonardis A and Stolkin R. 2018. Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognition, 76: 506-521 [DOI: 10.1016/j.patcog.2017.11.026]

Mishra O, Kapoor R and Tripathi M M. 2019. Human action recognition using modified bag of visual word based on spectral perception. International Journal of Image, Graphics and Signal Processing,

11(9)：34-43［DOI：10.5815/ijigsp.2019.09.04］

Mocanu I，Axinte D，Cramariuc O and Cramariuc B. 2018. Human activity recognition with convolution neural network using TIAGo robot//Proceedings of the 41st International Conference on Telecommunications and Signal Processing. Athens，Greece：IEEE：1-4［DOI：10.1109/TSP.2018.8441486］

Nazir S，Yousaf M H，Nebel J C and Velastin S A. 2018. A bag of expression framework for improved human action recognition. Pattern Recognition Letters，103：39-45［DOI：10.1016/j.patrec.2017.12.024］

Peng X J，Zou C Q，Qiao Y and Peng Q. 2014. Action recognition with stacked fisher vectors//Proceedings of the 13th European Conference on Computer Vision. Zurich，Switzerland：Springer：581-595［DOI：10.1007/978-3-319-10602-1_38］

Peng X J，Wang L M，Wang X X and Qiao Y. 2016. Bag of visual words and fusion methods for action recognition：comprehensive study and good practice. Computer Vision and Image Understanding，150：109-125［DOI：10.1016/j.cviu.2016.03.013］

Reddy K K and Shah M. 2013. Recognizing 50 human action categories of web videos. Machine Vision and Applications，24(5)：971-981［DOI：10.1007/s00138-012-0450-4］

Ren S Q，He K M，Girshick R and Sun J. 2017. Faster R-CNN：towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence，39(6)：1137-1149［DOI：10.1109/TPAMI.2016.2577031］

Russakovsky O，Deng J，Su H，Krause J，Satheesh S，Ma S，Huang Z H，Karpathy A，Khosla A，Bernstein M，Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision，115(3)：211-252［DOI：10.1007/s11263-015-0816-y］

Schydlo P，Rakovic M，Jamone L and Santos-Victor J. 2018. Anticipation in human-robot cooperation：a recurrent neural network approach for multiple action sequences prediction//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane，Australia：IEEE：1-6［DOI：10.1109/ICRA.2018.8460924］

Simonyan K and Zisserman A. 2014a. Two-stream convolutional networks for action recognition in videos//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge，USA：MIT Press：568-576

Simonyan K and Zisserman A. 2014b. Very deep convolutional networks for large-scale image recognition［EB/OL］.［2019-12-22］. https://arxiv.org/pdf/1409.1556.pdf

Tu Z G，Xie W，Qin Q Q，Poppe R，Veltkamp R C，Li B X and Yuan J S. 2018. Multi-stream CNN：learning representations based on human-related regions for action recognition. Pattern Recognition，79：32-43［DOI：10.1016/j.patcog.2018.01.020］

Wang H，Oneata D，Verbeek J and Schmid C. 2016a. A robust and efficient video representation for action recognition. International Journal of Computer Vision，119(3)：219-238［DOI：10.1007/s11263-015-0846-5］

Wang H and Schmid C. 2013. Action recognition with improved trajectories//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney，Australia：IEEE：3551-3558［DOI：10.1109/ICCV.2013.441］

Wang Y F，Song J，Wang L M，Van Gool L and Hilliges O. 2016b. Two-stream SR-CNNs for action recognition in videos//Proceedings of British Machine Vision Conference. York，UK：BMVA：108.1-108.12［DOI：10.5244/C.30.108］

Xiao J H，Cui X H and Li F. 2019. Human action recognition based on convolutional neural network and spatial pyramid representation. Journal of Visual Communication and Image Representation，71：1-10［DOI：10.1016/j.jvcir.2019.102722］

Yang Y Z，Li Y，Fermuller C and Aloimonos Y. 2015. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin，USA：AAAI：9286-9673

Zheng Y，Yao H X，Sun X S，Zhao S C and Porikli F. 2018. Distinctive action sketch for human action recognition. Signal Processing，144：323-332［DOI：10.1016/j.sigpro.2017.10.022］

Zhi Q and Cooperstock J R. 2012. Toward dynamic image mosaic generation with robustness to parallax. IEEE Transactions on Image Processing，21(1)：366-378［DOI：10.1109/tip.2011.2162743］

Zhu Y，Lan Z Z，Newsam S and Hauptmann A. 2018. Hidden two-stream convolutional networks for action recognition//Proceedings of the 14th Asian Conference on Computer Vision. Perth，Australia：Springer：363-378［DOI：10.1007/978-3-030-20893-6_23］

作者简介

马淼，1989 年生，女，讲师，主要研究方向为模式识别、图像处理、计算机视觉。
E-mail：mamiao@zstu.edu.cn

李贻斌，男，教授，主要研究方向为智能机器人、特种机器人、人机交互。E-mail：liyb@sdu.edu.cn

武宪青，男，讲师，主要研究方向为非线性控制、智能控制、信号处理。E-mail：wxq@zstu.edu.cn

高金凤，女，教授，主要研究方向为多智能体系统、多机器人协作、人机交互。E-mail：jfgao@zstu.edu.cn

潘海鹏，男，教授，主要研究方向为智能检测、控制、图像信息处理。E-mail：pan@zstu.edu.cn