文章编号：2095-2163(2021)10-0139-04 中图分类号：TP399 文献标志码：A

基于 SimplePose 优化算法的皮影保护技术

彭 然，刘爱丰，李斐逸，刘 扬，范冰冰，刘美奇
（四川农业大学 信息工程学院，四川 雅安 625014）

摘 要：针对传统皮影技艺的流失现象，本文将骨骼关键检测与传统皮影艺术结合。抛弃传统的网络结构复杂的 hourglass，cpn 模型，采用更加轻量、效果更好的 CNN 网络 SimplePose 模型，并对图像数据进行暗通道去雾、超微分图像超分辨率重建算法等操作进一步提取图像特征信息。不考虑算法本身而从实用出发，选取 ResNet-50 作为 Backbone 提高网络模型精度，在保证精度的同时减少参数量。将预处理视频逐帧切片，并映射为皮影图像，逐帧组合完成动作捕捉。

关键词：骨骼关键点；皮影；SimplePose

Protection of shadow puppet art based on SimplePose optimization algorithm

PENG Ran, LIU Aifeng, LI Feiyi, LIU Yang, FAN Bingbing, LIU Meiqi
(College of Information and Engineering, Sichuan Agricultural University, Ya'an Sichuan 625014, China)

【Abstract】Aiming at the loss of traditional shadow puppet skills, this article combines key bone detection with traditional shadow puppet art. In this paper, abandon the traditional hourglass, cpn model with complex network structure, adopt a lighter and better CNN network SimplePose model, and perform dark channel defogging, super-differential image super-resolution reconstruction algorithm and other operations on the image data to further extract the image characteristic information. Regardless of the algorithm itself but starting from practicality, ResNet-50 is selected as the Backbone to improve the accuracy of the network model and reduce the number of parameters while ensuring the accuracy. The pre-processed video is sliced frame by frame, mapped to shadow puppet images, and combined frame by frame to complete the motion capture.

【Key words】bone key points; shadow puppets; SimplePose

0 引 言

皮影戏是中国民间古老的传统艺术，老北京人都将其称为“驴皮影”。据史书记载，皮影戏始于西汉，兴于唐朝，盛于清代，元代时期传至西亚和欧洲，可谓历史悠久，源远流长。改革开放之后，皮影戏日渐式微，现在因受国家“非遗法”的保护，减缓衰萎的速度[1]。在以往对皮影和人的动作捕捉方法中，主要通过动画捕捉完成动作映射[2]，但成本较高且实现不易。在计算机视觉技术飞速发展的条件下，本文提出了一种使用卷积神经网络方法来捕捉人物模型动作，大大提升了动作映射的效率。通过此方法，极易将热门视频转化为皮影图像，引发人们对传统皮影技艺的兴趣，保护皮影艺术。

1 数据和方法

1.1 实验数据集

为了得到更好的模型效果，研究中采用 MPII 数据集对骨骼关键点进行提取，MPII 是用于评估人体姿势估计的数据集以及相关基准，拥有约 2.5 万张图像，并且包含超过 4 万名具有注释关节的人，该数据集利用人类活动的既定分类法系统化收集图像。表 1 包含有用于训练或验证的图像数量的信息。

表 1 数据集图片数量

Tab. 1 Number of picture sets

MPII 数据集	训练集	验证集	测试集
数据量	14 679	2 729	66 919

所使用的图像示例，如图 1 所示。图像以 JPG 格式进行存储。

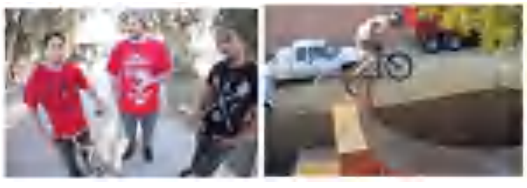

图 1 显示人体姿态图像示例

Fig. 1 An example of a human attitude image

作者简介：彭 然(2000-)，男，本科生，主要研究方向：计算机视觉、联邦学习；刘爱丰(2002-)，男，本科生，主要研究方向：深度学习、人工智能；李斐逸(2002-)，男，本科生，主要研究方向：人工智能、知识图谱；刘 扬(2002-)，男，本科生，主要研究方向：计算机视觉、机器学习；范冰冰(2001-)，女，本科生，主要研究方向：数据分析、数据清洗；刘美奇(2001-)，男，本科生，主要研究方向：机器学习、深度学习。

收稿日期：2021-08-06

1.2 研究处理方法

在本文的图像预处理中,着重探讨论述的是暗通道去雾。研究中,MPII 数据集每张图片均来自 YouTube 视频,图片的清晰度往往受外界环境影响,进而影响模型识别和判断结果。为了进一步提高模型的精度,更好地拟合模型状态,研究拟采用暗通道去雾算法[3]后得到的图像数据。

在图像的大多数局部区域,某些像素始终至少有一个值非常低的彩色通道,而此区域的最低光强度是很小的数字。暗通道的数学定义,对于任何输入图像 J, 暗通道可以表示为:

$$J^{dark}(x) = \min(\min J^c(y)),\ y \in \Omega(x), c \in \{r,g,b\}$$

当 J^c 表示彩色图像的每个通道时,$\Omega(x)$ 表示以像素 X 为中心的窗口暗通道先验理论, 由此可以得到:

$$J^{dark} \to 0 \tag{1}$$

计算机视觉中的雾图生成模型可写为:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{2}$$

其中, $I(x)$ 为无雾图像;$J(x)$ 为待恢复的原始无雾图像;A 为全球大气光分量;$t(x)$ 为透射率。由现有的 $I(x)$,即可求得 $J(x)$。

在此基础上,将其转换为如下公式:

$$\frac{I^c(x)}{A^c} = t(x)\frac{J^c(x)}{A^c} + (1 - t(x))$$

如果 C 表示 3 个通道,假设每个窗口中的传输是恒定的,并定义为 $\hat{t}(x)$,还给出了 A 值,则需要执行 2 个最小操作,即:

$$\min_c\left(\min\frac{I^c(y)}{A^c}\right) = \hat{t}(x)\min_c\left(\min\frac{J^c(y)}{A^c}\right) + (1 - t(x)) \tag{3}$$

上述 $\hat{J}^c(x)$ 是需要寻找的无雾图像,因此其暗通道应满足前一种情况:

$$J^{dark}(x) = \min(\min J^c(y)) = 0 \tag{4}$$

可以得到如下公式:

$$\min_c\left(\min\frac{J^c(y)}{A^c}\right) = 0 \tag{5}$$

引入要寻求的公式可以得出估计的 $\hat{t}(x)$ 结果为:

$$\hat{t}(x) = 1 - \min_c\left(\min\frac{J^c(y)}{A^c}\right) \tag{6}$$

图像增强效果如图 2 所示。

(a) 原图 (b) 取暗通道图 (c) 去雾后的图

图 2 暗通道去雾后展示图像

Fig. 2 The image is displayed after the dark channel is de-fogged

2 网络结构及数据再处理

2.1 网络结构

研究中,采用 CNN 模型 SimplePose,实现自上而下、即先找到人体,再判断关节点归属的人体骨骼关键点检测(Pose Estimation),网络结构在 ResNet 后加上几层反向卷积(Deconvolution)直接生成热力图,相比 Hourglass,CPN 等其他模型,使用 Deconvolution 替代了上采样结构。网络结构如图 3 所示。

图 3 SimplePose 网络结构图

Fig. 3 SimplePose network structure

这里值得一提的是,在 ResNet 的基础上,取最后残差模块输出特征层(命名 C5), SimplePose 采用 Deconv 扩大特征图的分辨率。Deconvolution 模型如图 4 所示。

图 4 Deconvolution 模型

Fig. 4 Deconvolution model

与其他经典算法性能对比[4]参见表 2。

表 2 与其他算法性能对比

Tab. 2 Performance comparison with other algorithms

Method	Backbone	Input size	AP	AP50	AP75	APm	APl	AR
CMU-Pose	-	-	61.8	84.9	67.5	57.1	68.2	66.5
Mask-RCNN	ResNet-50-FPN	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI	ResNet-101	353×257	64.9	85.5	71.3	62.3	70.0	69.7
CPN	ResNet-Inception	384×288	72.1	91.4	80.0	68.7	77.2	78.5
FAIR *	ResNeXt-101-FPN	-	69.2	90.4	77.0	64.9	76.3	75.2
G-RMI *	ResNet-152	353×257	71.0	87.9	77.7	69.0	75.2	75.8
oks *	-	-	72.0	90.3	79.7	67.6	78.4	77.1
bangbangren * + CPN+	ResNet-101	-	72.8	89.6	79.6	68.6	80.0	78.7
	ResNet-Inception	384×288	73.0	91.7	80.9	69.5	78.1	79.0
SimplePose	ResNet-152	384×288	73.7	91.9	81.1	70.3	80.0	79.0

根据文献[4]的实验数据，研究得到的仿真结果参见表 3，研究中又调整了输入图片尺寸。

表 3　输入图片大小对网络模型效果的影响对比

Tab. 3　Comparison of the influence of the input image size on the effect of the network model

Method	Backbone	Input size	#Deconv.Layers	Deconv. Kernel Size	*AP*
a	ResNet-50	256×192	3	4	70.4
b	ResNet-50	256×192	2	4	67.9
c	ResNet-50	256×192	3	2	70.1
d	ResNet-50	256×192	3	3	70.3
e	ResNet-101	256×192	3	4	71.4
f	ResNet-152	256×192	3	4	72.0
g	ResNet-50	128×96	3	4	60.6
h	ResNet-50	384×288	3	4	72.2

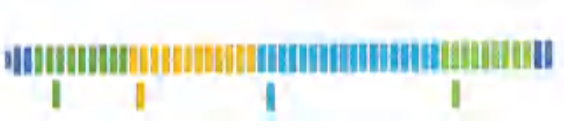

图 5　ResNet-50 网络结构图

Fig. 5　ResNet-50 network structure

2.2　数据再处理

将输入图像大小拓展至 384×288 像素。使用超微分图像超分辨率重建算法（SRResNet 算法）[5]，SRResNet 使用深度残差网络来构建超分重建模型，主要包含 2 部分：深度残差模型、子像素卷积模型。深度残差模型用来进行高效的特征提取，可以在一定程度上削弱图像噪点。子像素卷积模型主要用来放大图像尺寸。模型框架如图 6 所示。

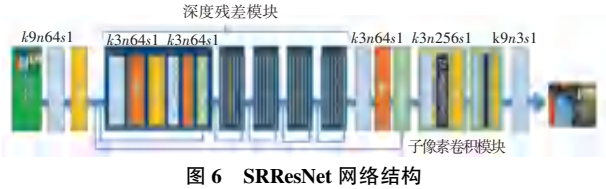

图 6　SRResNet 网络结构

Fig. 6　SRResNet network structure

图 6 中，k 表示卷积核大小，n 表示输出通道数，s 表示步长。除了深度残差模块和子像素卷积模块以外，在整个模型输入和输出部分均添加了一个卷积模块用于数据调整和增强。扩充图结果如图 7 所示。

图 7　扩充前后图像对比

Fig. 7　Image comparison before and after expansion

对人体骨骼关键点检测结果示例如图 8 所示。故研究中选取 ResNet-50 作为 Backbone，如图 5 所示。

图 8　骨骼关键点检测示例

Fig. 8　Example of bone key detection

3　在皮影戏上的运用

通过获取到的骨骼关键点，确定各个关节的位置，将人体关节与相应的皮影图片进行匹配，计算位置与旋转方向，达到人体活动与皮影运动同步，进行动作捕捉。从而促进中国传统技艺皮影戏的传承。

通过 2 个骨骼关键点可以确认肢体的长度和旋转角度，由于皮影面是二维平面，只需对应平面的 (x,y) 方向上的坐标，设 2 个点的坐标分别为 (x_1,y_1)、(x_2,y_2)，计算旋转角，将皮影素材图像按旋转角中心旋转，再计算 2 个关键点间的位移，得到映射点位置。并将各个素材图片映射到对应的肢体上，达到动作映射的效果，如图 9 所示。

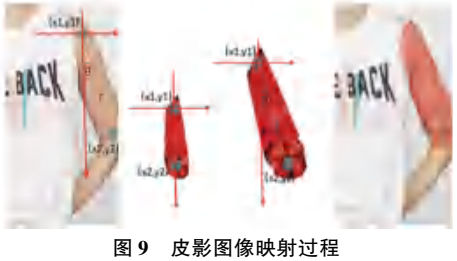

图 9　皮影图像映射过程

Fig. 9　Shadow image mapping process

将视频逐帧切割，并逐张映射为皮影图像，最后按帧聚合组装成皮影戏视频。

本次实验，选取单帧图像作为皮影映射后的展示图，如图 10 所示。

（下转第 146 页）