文章编号：2095-2163(2020)03-0032-08 中图分类号：TP391.41 文献标志码：A

基于顺序验证提取关键帧的行为识别

张 舟，吴克伟，高 扬
（合肥工业大学 计算机与信息学院，合肥 230601）

摘 要：人类行为识别作为视频分类中的重要问题，正成为计算机视觉中的热门话题。由于视频信息较多，有的视频冗余信息过量，判别性帧较少，因此如何无监督地提取关键帧对于行为识别至关重要。为此，本文提出了一种新的基于顺序验证的关键帧提取方法，并将其应用到行为识别中。首先，本文定义了一种顺序验证的模块，验证局部区间中帧的顺序，学习局部区间中帧的关键性描述，接着将其整合得到整段视频中每一帧的关键性描述；其次，根据学习到的视频帧关键性描述提取关键帧；最后通过实验讨论分析提取多少关键帧对行为识别最有利。实验结果表明，本文的方法在 UCF-101 上可以达到 95.40%，在 HMDB51 上可以达到 68.80%，均优于当前的一些先进的方法。

关键词：行为识别；关键帧提取；顺序验证；关键性描述

Action recognition based on key frame extraction using order verification

ZHANG Zhou, WU Kewei, GAO Yang
(School of Computer and Information, Hefei University of Technology, Hefei 230601, China)

【Abstract】 As an important issue in video classification, human action recognition is becoming a hot topic in computer vision. Since there are many video information, some videos have redundant information and few discriminative frames, so how to extract key frames unsupervised is very important for action recognition. To this end, the paper proposes a new key frame extraction method based on order verification and apply it to action recognition. First, this paper defines an order verification module that verifies the order of frames in a local interval, learns the key description of the frames in the local interval, and then integrates them to obtain the key description of each frame in the entire video; Second, key frames are extracted based on the learned key descriptions of the video frames; Finally, the paper discusses experimentally how many key frames are extracted to be most beneficial for action recognition. Experimental results show that the proposed method can reach 95.40% on UCF-101 and 68.80% on HMDB51, which are all better than some current advanced methods.

【Key words】 action recognition; key frame extraction; order verification; key description

0 引 言

视频中的人体行为识别是计算机视觉领域的一项既基础又具有挑战性的任务，最近几年正被广泛应用于视频监控、人机交互、医疗看护等领域[1]。这个任务是指从视频序列中提取相关的视觉信息，并用合适的方式表达出来，然后通过对视觉信息的解释来分析和识别人类的行为模式。真实的视频大多以人类活动为背景，在视频某些时间段里背景比较复杂，很难准确、鲁棒地识别人类行为，因此行为识别仍是一个复杂的问题。

现有的深度学习模型，将行为识别任务视为多分类问题。其早期研究关注于利用卷积神经网络(CNN)来学习视频中行为的深度表达，包括双流 CNN 模型[2]，隐双流 CNN 模型[3]，以及 3D-CNN 模型[4]。卷积神经网络擅长于捕获场景的空间信息，然而其对时序信息的捕获能力不强。现有深度学习模型通常使用循环神经网络(RNN)，尤其是长短期记忆网络(LSTM)模型来描述行为中时序信息。现有行为识别的难点在于，目标动作仅仅占长视频中的一小部分，同时运动目标被大量的背景信息干扰，因此，从长视频中提取行为发生的有效信息，成为行为识别的关键问题。

针对现有方法无法有效区分视频中时序背景混杂信息，导致行为识别准确率和效率不高的情况，研究发掘了一种基于长视频序列顺序验证的新的关键帧提取方法，并将这种方法应用到行为识别中去。在此方法中，通过抑制视频中的低质量时序信息，学习到具有辨别性的视频帧的表示，提高行为表达的判决能力，从而实现可靠的行为识别。综上所述，本次研究做出以下贡献：

(1)本文提出了一种新的基于顺序验证提取关键帧的行为识别方法。其中，这种关键帧机制用于

作者简介：张 舟(1995-)，男，硕士研究生，主要研究方向：计算机视觉；吴克伟(1984-)，男，副研究员，主要研究方向：计算机视觉；高 扬(1996-)，男，硕士研究生，主要研究方向：计算机视觉。

通讯作者：张 舟 Email：18098789799@163.com

收稿日期：2019-12-16

去除低质量背景复杂的冗余帧，然后将这种关键帧机制应用到行为识别任务中。

（2）本文设计了一种顺序验证的方法来学习视频帧的关键性描述。首先验证局部区间中帧之间的顺序关系，获取局部区间中帧的关键性描述；然后以某种方式结合各阶段局部区间中帧的关键性描述，得到整段视频中每一帧的关键性描述。

（3）本文进一步将关键帧提取应用到了行为识别上，并在 UCF101 和 HMDB51 这 2 个公认的数据集上进行实验验证。实验结果表明，在 UCF101 上提取 12 帧关键帧表现最好，识别精度为 95.40%，在 HMDB51 上提取 10 帧关键帧表现最好，识别精度为 68.80%，均优于目前大部分先进的方法。

1　相关工作

视频相比图像来说信息更加丰富，但是一个视频序列中冗余信息太多，如何高效准确地提取关键帧的信息对于很多任务都是至关重要的。与此同时人类行为识别是计算机视觉领域一个长期存在的课题，也是当今一个研究热点。在这部分，分别介绍了关键帧提取和行为识别两方面的相关工作。

（1）关键帧提取。许多早期的关键帧提取方法依赖于使用基于管道的分割，此类方法通常提取光流和 SIFT 特征。较早的方法[5]通过视频的光流检测了连续帧之间的相似性的局部最小变化。之后的方法通过在特征提取中使用关键点检测[6-7]改进了这一点，后者通过 SIFT 描述符提取局部特征，并汇总了关键点以实现视频中的关键帧提取。但是，所有这些方法都具有以下缺点：当相同的内容再次出现在视频中时，就可能会提取相似的关键帧。另一类方法是将视频帧的特征（如 HS 颜色直方图）聚类成组。这些方法通过从每个组中检测有代表性的帧来确定视频中的关键帧。Zhuang 等人[8]提出了一种基于视觉内容和运动分析的关键帧非监督聚类方法。Vázquez 等人[9]提出了一种基于频谱聚类的关键帧检测方法，该方法构建了一个图来捕获图像视频序列中的特征局部性，而不是依靠由 2 个图像之间共享的特征所计算出的相似性度量。最后由于 CNN 在图像分类中的流行，已将 CNN 引入视频的关键帧提取中。Mahasseni 等人[10]首先将生成对抗网络（GAN）应用于视频中的关键帧提取。

（2）行为识别方法。同时，CNN 在图像分析任务中深度特征提取的成功，为视频中行为分类的研究提供了灵感。CNN 侧重空间模式的提取，可以有效增强行为特征在空间域上的表现能力，比如在 ImageNet[11] 数据集上预训练的 Vggnet[12]、GoogleNet[13]和 ResNet[14]，并将其用作特征提取器。此外，Zhu 等人[3]提出了一种新型的 Hidden Two-stream CNN 架构，隐式地捕获相邻帧之间的运动信息。Wang 等人[15]提出了一种新的架构，称为外观-关系网络（ARTNet），以端到端的方式学习视频表示，ARTNet 是通过堆叠多个 SMART 块来构建的。Shou 等人[16]提出了一种轻量级的生成器网络，该网络减少了运动矢量中的噪声，捕获了精细的运动细节，实现了一种更具鉴别性的运动线索（DMC）表示。但是由于 CNN 对时序信息的捕获能力不强，而 RNN 具有学习帧之间时序关系的强大能力，尤其是 LSTM 网络由于其灵活的门机制，可以避免在反向传播过程中梯度消失或梯度爆炸。Li 等人[17]提出了一个新颖的框架，通过结合 CNN 和 LSTM 来学习视频中的时序动态特征，从而达到增强行为识别的效果。Ng 等人[18]通过实验证明，相较于传统的双流方法[2]，加入 LSTM 整合时序信息可以显著提高行为识别的准确率。

（3）关键帧提取用于行为识别。视频并非每一帧都有对行为识别有利的信息，因此去除冗余帧，将关键帧机制加入行为识别任务有着重大的意义。Wang 等人[19]提出了一种从视频序列中提取人类动作识别关键帧的新方法，主要利用研究提出的一种自适应加权亲和传播算法（SWAP），以提取关键帧，最后结合 SVM 进行行为识别。但是这种方法对识别精度贡献并不大，只是改善了识别速度。Zhou 等人[20]提出一种实时的行为识别方法，通过这种从视频帧的时间窗口中检测关键帧的新算法来提高识别速度，再采用隐马尔可夫模型（HMM）来分析检测到的关键帧的时间关系，从而保证识别的准确性。同样，为了弥补高斯混合隐马尔可夫模型（GMM-HMM）需要定义高斯混合模型（GMM）和隐马尔可夫模型（HMM）分类的数量，从而引起的识别速度下降，Li 等人[21]提出了一种基于关键帧的 GMM-HMM 运动识别方法，使用最小重建误差方法来确定关键帧的数量，从而减少 GMM 和 HMM 分类的数量提高识别速率。Zhao 等人[22]提出一种新的基于关键帧提取和多特征融合技术的行为识别方法，既利用关键帧机制解决了数据冗余的问题，又通过多特征融合不同流的信息，提高了识别精度。Zhu 等人[23]通过挖掘视频中关键帧所在视频段来提高识别正确率。Kar 等人[24]采用含有时空网络和 MIL 框架的双流 CNN 来检测视频中得分较高的关键帧，

进而应用于行为识别。

受到文献[25-26]采用顺序验证来进行行为识别的启发,且目前没有基于顺序验证来学习关键帧的方法,本文提出一种顺序验证的方法,提取视频中的关键帧,去除冗余信息,进而再将这种新的关键帧提取方法用于视频中的行为识别,实验结果表明本文的方法取得了较好的识别正确率。

2 模型框架

在本节中,首先对所提出的方法给出简要论述,然后将本文方法的每个部分进行详细阐明。这里,以 UCF101 数据集为例,研究得到的本文模型的视频整体序列化处理过程如图 1 所示。相应地,行为识别的网络架构可以分为以下 4 个模块:提取 CNN 特征(2.1 节);顺序验证(2.2 节);学习关键帧(2.3 节);最终的行为识别(2.4 节)。首先,采用 CNN 是因其在图像特征提取方面的成功应用;其次,设计了一个局部的顺序验证模型,通过对局部顺序验证结果的分析,计算局部区间中帧的关键性描述;再者,将局部区间中帧的关键性描述相结合,形成整段视频中帧的关键性描述,并进行关键帧提取;讨论截取关键帧的数目,提取出相应数目的关键帧;最后,设计了一种新的基于顺序验证的关键帧提取的行为识别框架来识别人类的行为。

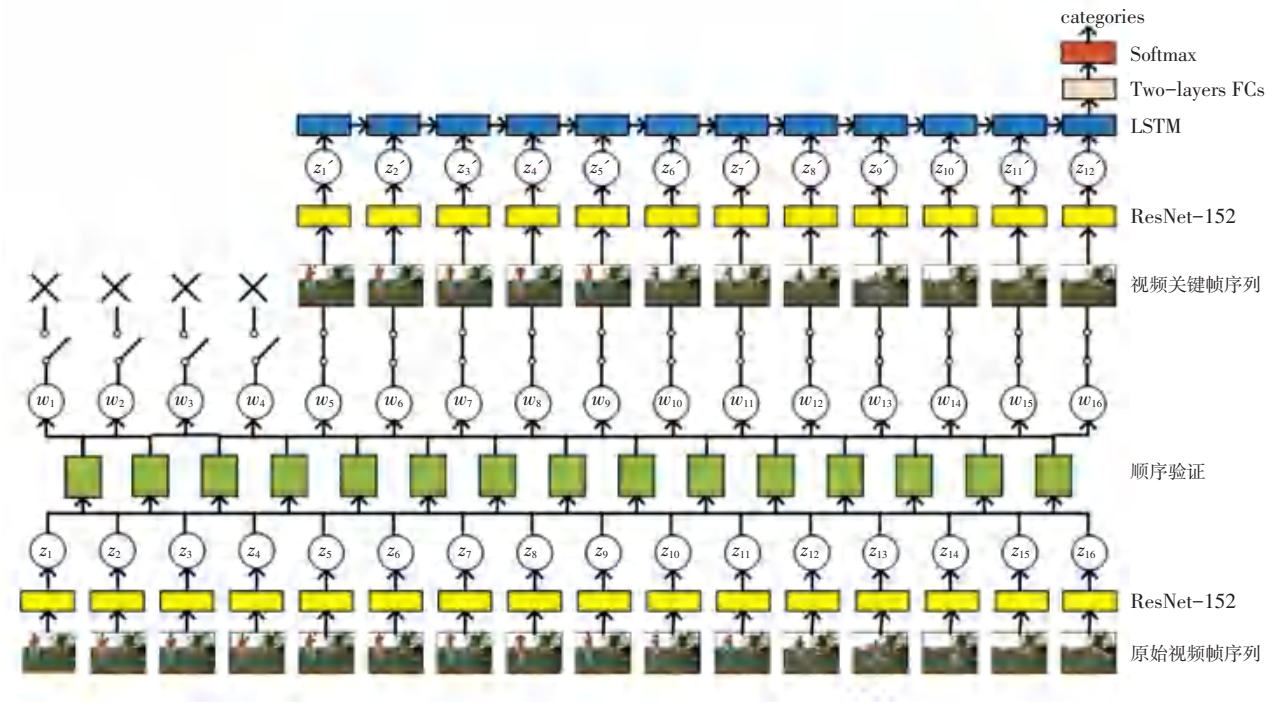


图 1 本文模型的视频整体序列化处理过程

Fig. 1 The overall video serialization process of the model in this paper

本文模型的主要创新价值在于:

(1)提出了一种新的基于顺序验证的关键帧提取方法,并将其用于视频的行为识别中。

(2)为了有效估计视频帧的关键性,设计了一个顺序验证模块来验证局部视频段中帧之间的顺序。将局部视频段的长度设置为 2 个连续视频帧,通过对局部顺序验证结果的分析,计算局部区间中帧的关键性描述;再者,将每段视频内局部区间中帧的关键性描述相结合,形成整段视频中每一帧的关键性描述,并排序。

(3)为了达到最佳的识别效果,进行了多组对比实验分析提取关键帧的数目,最终确定在 UCF101 上每段视频提取 12 个视频帧,在 HMDB51 上每段视频提取 10 个视频帧。

2.1 特征提取

识别视频中的行为往往不需要通过视频中的所有帧,只需选择一些帧组成序列来代表这个视频。因此将一个有 L 帧的视频分成 $16 = L/\alpha$ 个非重叠的单元,每个单元包含 α 个连续的帧。然后在每个单元中选择第一帧,组合形成帧序列 $V = \{v_t\}$,$(t = 1, 2, \cdots, 16)$。研究中提取这些视频帧的外观特征用于行为表达,为此,本文使用在 ImageNet 数据集上

预训练好的 ResNet-152 模型，对已经重新调节大小为 224×224 的 RGB 图像序列进行预处理，对于第 t 帧提取输入最后一层全连接层之前的结果作为最终特征：z_t，在此基础上，通过时序 SVM 网络对特征序列进行建模。

2.2　顺序验证

所提出的顺序验证模块如图 2 所示。由图 2 可知，该模块具有 3 个主要组成部分：二元组采样；使用时序 SVM 进行局部区间顺序验证得到局部区间内视频帧的关键性描述；将局部区间内视频帧的关键性描述整合到整段视频中，得到每个视频帧最终的关键性描述。对此可做阐释分述如下 。

(1)二元组采样。采用了 SVM 在顺序验证模块上训练网络，因此如何采样正负例是一个关键挑战。为了解决这个问题，将每个视频帧序列中相邻两帧作为一个二元组 $(v_a,v_b)(a<b)$，其中正例为 (v_a,v_b)，负例为 (v_b,v_a)。一个视频的帧序列长度是 16，此处将其分成 15 个二元组 $(v_1,v_2),(v_2,v_3),\cdots,(v_{15},v_{16})$，如前文所述，就有 15 组正负例，因此本次研究构建了 15 个分类器训练这些正负例。

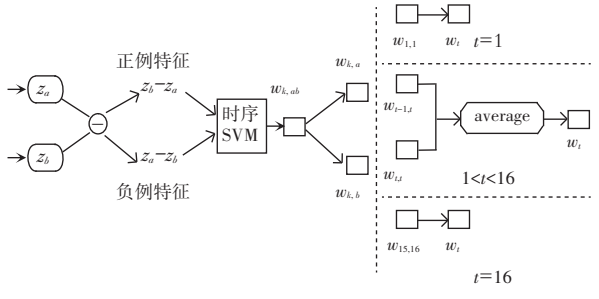


图 2　顺序验证模块

Fig. 2　Order verification module

(2)局部顺序验证。这部分验证的是局部区间中帧的顺序是否为正确的时序，因此可以被看作一个分类任务。考虑到要解决的是时序验证问题，文中选用了 SVM 来进行分类，将这个分类器称为时序 SVM。关于如何构建正负例特征，研究使用的是一个简单的减法，如图 2(a)所示，将组合的成对特征相减，得到的是：z_b-z_a 和 $z_a-z_b(a<b)$，其中 z_b-z_a 表示正例特征，z_a-z_b 表示负例特征。具体是将每个视频的 15 组待验证的正例输入训练好的时序 SVM 中得到属于正类的得分，再将这个得分视为每个局部区间帧对 (v_a,v_b) 的关键性描述 $w_{k,ab}$(k 表示第 k 个二元组)，其数学公式可表示为：

$$w_{k,ab}=SVM(z_a,z_b). \tag{1}$$

本文规定，对于帧 v_a 和 v_b，如图 2 虚线左侧所示，在这个二元组中的局部关键性描述分别为 $w_{k,a}=w_{k,ab}$ 和 $w_{k,b}=w_{k,ab}$。

(3)整段视频的关键性描述。为了学习整段视频中的每个选定帧(16 帧)的关键性，研究中将帧的局部关键性描述整合得到整段视频中帧的关键性描述如图 2(b)所示，并将用到如下数学公式：

$$w_t=\begin{cases} w_{1,1}, & t=1;\\ \dfrac{(w_{t-1,t}+w_{t,t})}{2}, & 1<t<16;\\ w_{15,16}, & t=16. \end{cases} \tag{2}$$

其中，w_t 表示视频序列的第 t 帧最终的关键性描述。

2.3　关键帧提取

针对视频中选定的每一帧，关键性描述在 2.2 节已经计算得出，如何进行关键帧的提取从而得到最佳识别效果是需经实验讨论的。首先原始的视频序列为 $V=\{v_t\},(t=1,2,\cdots,16)$，视频帧序列对应的关键性描述为 $W=\{w_t\}$。则关键帧提取的过程如下：

$$(W',V')=top(W,k). \tag{3}$$

其中，$top(W,k)$ 表示从 W 中由高到低取 k 个数据组成新的序列 W'，并取这 k 个数据对应的 k 帧关键帧的集合 V'。

2.4　行为识别

本文的顺序验证模块可以学习视频中帧的关键性，将关键性低的冗余帧去除，即实现了视频帧序列的关键帧提取。进而又将关键帧提取应用于视频的行为识别任务中。上一节中，已提取关键帧组成序列 V'，在这里，将提取序列中关键帧的 ResNet-152 特征得到对应的特征序列 $Z'=\{z'_1,z'_2,\cdots,z'_k\}$，采用了 LSTM 整合新的时序关系，有效地增强了模型对行为表达的能力，达到更好的行为识别效果。运算时需参考的数学公式可表示为：

$$\begin{cases} f_t=\sigma(\theta_F[h_{t-1},z'_t]+b_F);\\ i_t=\sigma(\theta_i[h_{t-1},z'_t]+b_i);\\ g_t=\tanh(\theta_c[h_{t-1},z'_t]+b_g);\\ o_t=\sigma(\theta_o[h_{t-1},z'_t]+b_o);\\ C_t=f_t\circ C_{t-1}+i_t\circ g_t;\\ h_t=o_t\circ\tanh(C_t). \end{cases} \quad t\in[1,k] \tag{4}$$

其中，θ 和 b 表示 LSTM 参数；f_t 是忘记门；i_t 是输入门；o_t 是输出门；g_t 表示 t 时刻候选记忆单元状

态；C_t 和 h_t 表示 t 时刻记忆单元状态和隐单元状态；z'_t 表示 t 时刻的输入特征；$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 表示 sigmoid 和 tanh 激活函数；“∘”表示哈达马积。

LSTM 模型的核心就是忘记门和输入门，忘记门会根据当前的输入 z'_t、上一时刻状态 C_{t-1} 和上一时刻输出 h_{t-1} 共同决定哪一部分记忆需要被遗忘。输入门会根据 z_t，C_{t-1} 和 h_{t-1} 决定哪些部分将进入当前时刻的状态 C_t。LSTM 结构在计算得到新的状态 C_t 后，会通过输出门根据最新的状态 C_t、上一时刻的输出 h_{t-1} 和当前的输入 z'_t 来决定该时刻的输出 h_t。

其中，k 在 UCF101 上为 12，在 HMDB51 上是 10，后不赘述。经过关键帧提取之后的 LSTM，将其最后一个隐藏单元的输出 h_k 输入两层全连接层从而得到一个 1×101 的向量 $P=\varphi(h_k)$，用来表示输入视频在每个类别上的得分。最后，将 P 经过 softmax 处理得到最终的分类类别。

2.5 损失函数

本文采用的是交叉熵损失函数，用来评估当前训练得到的概率分布与真实分布的差异情况，适用于分类学习。本文的样本损失函数如下：

$$L=-\sum_{g=1}^{j} y_g \log P_g. \tag{5}$$

其中，y_g 和 P_g 分别为输入视频帧属于真实类别编号 g 的真实概率值和模型预测概率值；j 为类别种类数，得到的损失 L 反向传播以优化整个框架。

3 实验

本节中，首先对数据集做了整体概述，然后阐述本文的实验过程及评价标准，最后对实验结果进行说明及讨论。

3.1 数据集

本文方法所用的数据集为 UCF101[27] 和 HMDB51[28]。UCF101 数据集包含 13 320 个视频，分为 101 个类别，使用 9 990 个视频用于训练，剩下的 3 330 个视频用于测试。UCF101 数据集在行为类别方面提供了多样性，并且在目标外观和姿态、背景杂乱、光照条件等方面存在巨大的变化。

HMDB51 数据集中包含 6 849 个视频，共 51 个行为类别，本文选取 4 794 个视频用于训练，其余的 2 055个视频用于测试。HMDB51 数据集在物体外观和人物姿态等方面变化多样，具有行为识别研究的挑战性。

3.2 实验设计及评价标准

为了准备训练特征集合，首先，依次提取各视频的 RGB 视频帧，并将分辨率重新调整为 224×224。其次，使用 ImageNet 数据集上预训练的 ResNet 模型，提取外观特征，具体来说，本文取 ResNet 输入最后一层全连接层之前的特征作为 LSTM 模型的输入特征，该特征的大小为 1×2 048，即 LSTM 模型的隐状态和记忆状态的维度为 2 048。

本文实验所采用的 PC 机配置为 Intel Core i7-5960X、CPU 3 GHz×8 cores RAM 8 GB、图像显卡为 2 张 NVIDIA GeForce GTX 1080 Ti、Linux16.04 操作系统。深度学习框架为 Pytorch[29]。训练时，使用 Adam 算法，迭代次数为 50，批处理大小为 128，学习率初始化为 10^{-3}。

本文采用识别正确率，作为行为识别的评价标准，即统计一个类别中的所有视频的预测标记被识别为真实标记的数值，与预测视频总数的比值，作为该类别的识别正确率；最后使用所有类别正确率的均值，作为本文方法的识别正确率。

3.3 实验结果及分析

本文与当前比较先进的行为识别方法进行了对比，根据加入关键帧机制与否，可以分为以下 2 组：

(1)带有关键帧机制的模型，包括：传统的双流 CNN 模型 Two-stream mode[2]，使用 CNN 进行还原分辨率隐式运动预测的模型 Hidden Two-Stream[13]，双流通道的时间池化模型 Beyond Short Snippets Models[5]，轻量级的生成器网络 DMC-Net[16]，通过堆叠多个可以同时对外观和时间关系进行建模的 SMART 模块的 ARTNet[15] 模型。

(2)带有关键帧机制的模型，包括挖掘识别关键帧所在视频段进行行为识别的模型 Key Volume Mining[23]，使用深度网络获得的特征经过 Adaptive Pooling 的方法进行关键帧提取的 AdaScan[24] 行为识别模型。

不同方法的识别性能对比见表 1。由表 1 分析可知，与当前一些优秀方法相比，本文方法所得到的识别正确率更高。相比于不带关键帧机制的方法而言，本文将关键帧提取加入到行为识别中去，在识别的过程中，因为减少了冗余帧，大大提升了识别的效率和准确率；相比于带关键帧的模型，本文先是精确定位到具有判别性的帧，相较于 Key Volume Mining 方法定位到关键帧所在视频段更为精确，再者较 AdaScan 采用 pooling 的方式对视频帧的关键性进行判定从而在测试过程中舍去冗余帧，本文既考虑前后帧之间的时序关系采用一种新的方法来判别帧的关键性，又通过 LSTM 的结构将视频中新的时序关

系加以整合，显著提升了识别正确率。为了更进一步证明本文加入关键帧机制对行为识别贡献显著，本文在 UCF101 和 HMDB51 两个数据集上进行了消融实验，结果见表 2。

表 1　不同方法识别性能对比

Tab. 1　Comparison of recognition performance of different methods　%

方法	Two-stream model[2] (2014)	Beyond Short Snippets Models[18] (2015)	Hidden Two-Stream[3] (2017)	DMC-Net[16] (2019)	ARTNet[15] (2017)	Key Volume Mining[23] (2016)	AdaScan[24] (2017)	本文方法
UCF101	88.00	88.60	90.30	90.90	93.50	93.10	93.20	**95.40**
HMDB51	59.40	NA	60.50	62.80	67.60	63.30	66.90	**68.80**

表 2　关键帧因素对实验结果的影响

Tab. 2　Influence of key frame factors on experimental results　%

	消融模型	UCF101	HMDB51
ResNet-152	LSTM	91.20	63.70
	Key frame(15)+LSTM	92.10	64.50
	Key frame(14)+LSTM	93.70	65.10
	Key frame(13)+LSTM	94.30	65.80
	Key frame(12)+LSTM	**95.40**	66.70
	Key frame(11)+LSTM	92.70	67.30
	Key frame(10)+LSTM	90.50	**68.80**
	Key frame(9)+LSTM	88.30	66.30

通过表 2 可以看到：

(1)本文提出的关键帧机制在 UCF101 数据集上，随着从初始的 16 帧按照关键性描述由低到高逐一去除冗余帧，识别正确率一路上升，直到去除 4 帧时达到最高的识别正确率 95.40%，此后继续去除则造成识别正确率下降，所以提取 12 个关键帧能达到最佳的识别效果；同理，在 HMDB51 数据集上，提取 10 个关键帧能达到最佳的识别效果。

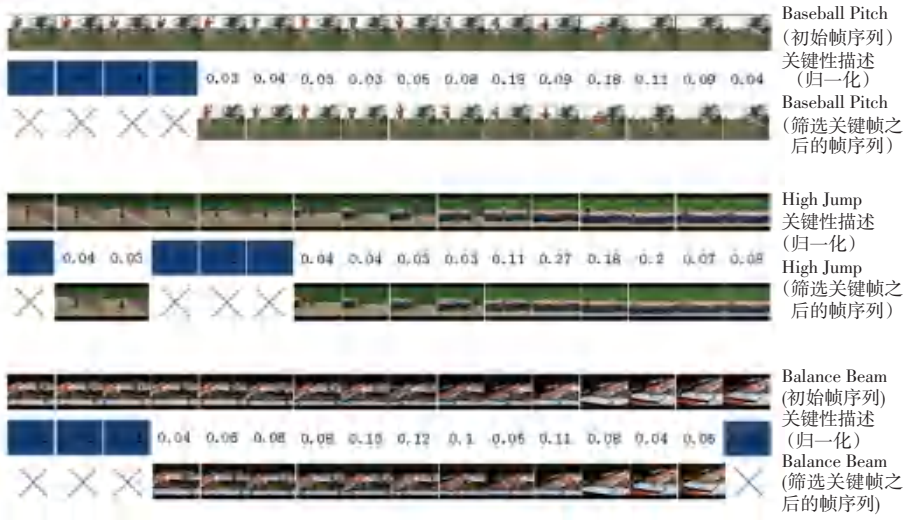
(2)在 UCF101 和 HMDB51 两个数据集中，本文提出的加入关键帧机制的行为识别模型的行为识别正确率全面优于无关键帧机制的行为识别模型，UCF101 上提升了 4.2%，HMDB51 上提升了 5.1%。由此说明本文所提出的关键帧机制可以有效地提取有辨别性的特征，从而可以增强行为的表达。

2.2 节中学习到了视频中每一帧的关键性描述，接着就是要进行关键帧提取，本次研究用实例图来表现关键帧提取的结果，如图 3 所示。在 UCF101 和 HMDB51 数据集中，分别随机选取代表 3 种行为的视频，观察其帧序列中每一帧的关键性描述，进而了解提取关键帧的过程。图 3(a)上、中、下三组分别表示的行为是“Baseball Pitch”、“High Jump”和“Balance Beam”，图 3(b)上、中、下三组分别表示的行为是“Throw”、“Kick Ball”和“Golf ”。每组图片中，第一行表示原始视频帧序列；第二行表示视频帧对应的归一化之后的关键性描述，数字越大，代表这一帧关键程度越高；第三行尝试去除关键性最低四帧后重新组合的视频帧序列，即提取出的关键帧序列。

分析图 3 可以看出，前后两帧几乎没有变化的动作帧，关键程度都比较低，进而本文的模型会抓取对判别该行为贡献较大的帧、即关键程度较高的帧，更加体现本文模型的判别能力。

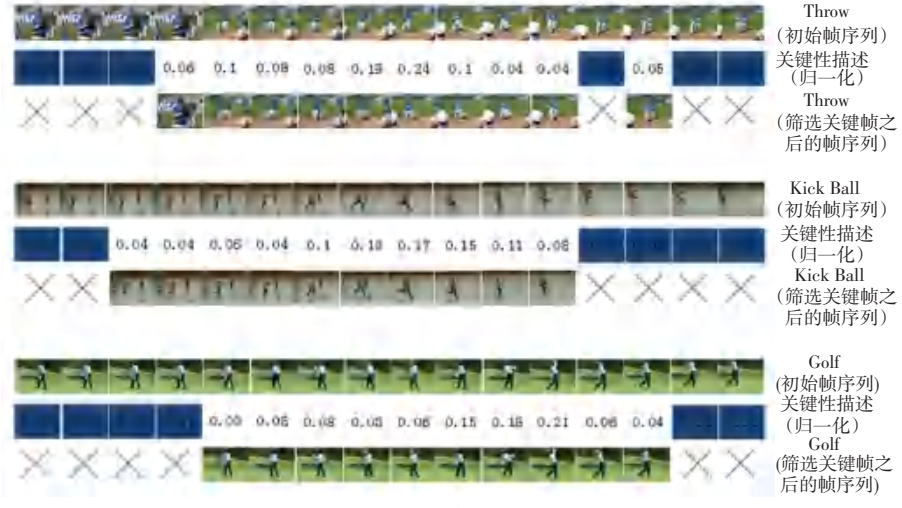
4　结束语

针对现有基于视频整体结构建模的行为识别方法，无法有效区分关键帧与冗余帧，造成行为表达效率低下，行为识别准确率不高的问题，本文提出了一种基于顺序验证提取关键帧的行为识别模型。通过在 UCF101 和 HMDB51 两个公认数据集上进行实验验证，可以证明本文的顺序验证模块能够识别关键帧，提高了行为表达的判决能力。在 UCF101 和 HMDB51 两个公认数据集上进行实验验证，与现有多种优秀的行为识别方法进行比较。实验结果表明，本文方法优于现有大部分行为识别方法。未来可以预期的是，本文的方法可以应用于更加复杂的视频场景中，如大型监控场景下的视频理解，异常检测等，将有助于维护公共安全等领域。



(a) 实例图 1

(a) Example 1



(b) 实例图 2

(b) Example 2

图 3 关键帧提取

Fig. 3 Key frame extraction

参考文献

[1] POPPE R. A survey on vision-based human action recognition[J]. Image and Vision Computing, 2010, 28(6): 976.

[2] SIMONYAN K , ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Computational Linguistics, 2014, 1(4):568.

[3] ZHU Yi , LAN Zhenzhong, NEWSAM S , et al. Hidden two-stream convolutional networks for action recognition[J]. arXiv preprint arXiv:1704.00389, 2017.

[4] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221.

[5] KULHARE S, SAH S, PILLAI S, et al. Key frame extraction for salient activity recognition [C]//2016 23rd International Conference on Pattern Recognition (ICPR). Cancun, Mexico: IEEE, 2016: 835.

[6] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91.

[7] GUAN Genliang, WANG Zhiyong, LU Shiyang, et al. Keypoint-based keyframe selection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 23(4): 729.

[8] ZHUANG Y, RUI Y, HUANG T S, et al. Adaptive key frame extraction using unsupervised clustering [C]//Proceedings of International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269). Washington DC,USA:IEEE, 1998, 1: 866.

[9] VÁZQUEZ-MARTÍN R, BANDERA A. Spatio-temporal feature-based keyframe detection from video shots using spectral clustering[J]. Pattern Recognition Letters, 2013, 34(7): 770.

[10] MAHASSENI B, LAM M, TODOROVIC S. Unsupervised video summarization with adversarial LSTM networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,HI:IEEE, 2017: 202.

[11] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-

scale hierarchical image database [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.

[13] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: CVPR, 2015:1.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE Computer Society, 2016:770.

[15] WANG Limin, LI Wei, LI Wen, et al. Appearance-and-relation networks for video classification [J]. arXiv preprint arXiv: 1711.09125, 2017.

[16] SHOU Z, LIN X, KALANTIDIS Y, et al. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 1268.

[17] LI Qing, QIU Zhaofan, YAO Ting, et al. Action recognition by learning deep multi-granular spatio-temporal video representation [C]//Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. New York, USA: ACM, 2016:159.

[18] NG Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: USA, 2015:4694.

[19] WANG Yongxiong, SUN Shuxin, DING Xueming. A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition [J]. Journal of Visual Communication and Image Representation, 2015, 33(C):1993.

[20] ZHOU Ling, NAGAHASHI H. Real-time action recognition based on key frame detection [C]//Proceedings of the 9th International Conference on Machine Learning and Computing. Singapore: ACM, 2017: 272.

[21] LI Jinhong, LEI Tingsheng, ZHANG Fengquan. An Gaussian-mixture hidden Markov models for action recognition based on key frame[C]//2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Beijing, China: IEEE, 2018: 1.

[22] ZHAO Y, GAO L, HE D, et al. Multi-feature fusion action recognition based on key frames[C]//2019 Seventh International Conference on Advanced Cloud and Big Data (CBD). Suzhou, China: IEEE, 2019: 279.

[23] ZHU W, HU J, SUN G, et al. A key volume mining deep framework for action recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2016: 1991.

[24] KAR A, RAI N, SIKKA K, et al. Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 3376.

[25] MISRA I, ZITNICK C L, HEBERT M. Shuffle and learn: Unsupervised learning using temporal order verification[C]//14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: dblp, 2016:524.

[26] LEE H Y, HUANG J B, SINGH M K, et al. Unsupervised representation learning by sorting sequences [C]//IEEE International Conference on ComputerVision (ICCV). Venice, Italy: IEEE, 2017:1.

[27] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[28] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition [C]//2011 IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011: 2556.

[29] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library [C]//33rd Conference on Neural Information Processing System (NeurIPS 2019). Vancouver, Canada: NIPS, 2019: 8024.