

AugmentGest: 随机数据裁剪增强能否提升手势识别性能？

Nada Aboudeshish

Dmitry Ignatov

Radu Timofte

Computer Vision Lab, CAIDAS, University of Würzburg, Germany

Abstract

数据增强是一种在深度学习中至关重要的技术，尤其适用于那些具有有限数据集多样性的任务，例如基于骨架的数据集。本文提出了一个综合的数据增强框架，该框架整合了几何变换、随机裁剪、旋转、缩放和基于强度的变换、亮度和对比度调整，以模拟现实世界的多样性。随机裁剪确保了时空完整性的保持，同时应对了视点偏差和遮挡等挑战。增强管道为数据集中的每个样本生成三个增强版本，从而使数据集的大小增加到四倍，并丰富了手势表示的多样性。所提出的增强策略在三种模型上进行评估：多流 e2eET 模型、基于点云的手势识别 (HGR) 的 FPPR 模型和 DD-Network 模型。实验在包括 DHG14/28、SHREC' 17 和 JHMDB 的基准数据集上进行。e2eET 模型是 DHG14/28 和 SHREC' 17 上手势识别的最新技术。FPPR-PCD 模型是 SHREC' 17 上表现第二好的模型，在基于点云的手势识别中表现出色。DD-Net 是一种面向基于骨架的动作识别的轻量高效架构，在 SHREC' 17 和人类运动数据库 (JHMDB) 上进行评估。结果表明，所提出的增强策略在各种数据集和架构上显著提高了模型的泛化能力和鲁棒性。该框架不仅在所有三个评估模型上确立了最新技术结果，还提供了可扩展的解决方案，以推动手势识别和动作识别在现实场景中的应用。该框架在 <https://github.com/NadaAbodeshish/Random-Cropping-augmentation-HGR> 上可获取。

索引词—手势识别，数据增强，数据级融合，骨架数据

1. 介绍

手势识别已成为人机交互的基石，为游戏、虚拟和增强现实、机器人技术以及辅助技术等应用打开了新的可能性。作为一种直观的交流方法，手势识别在提升用户体验和实现与技术的自然交互中发挥了关键作用。尽管在深度学习和计算机视觉领域取得了显著进展，开发稳健且具有普遍适应性的手势识别系统仍然具有挑战性。数据集多样性有限、视角变化以及真实世界的复杂性等因素持续阻碍着手势识别模型的性能。数据增强已被证明是一种强大的技术，通过人工增加训练数据集的多样性和规模，从而增强模型的稳健性和泛化

能力。几何变换和基于强度的调整等技术在基于图像的任务中已显示出显著的前景。然而，它们在改善基于骨架的手势识别方面的潜力尚未被充分探索，在这类识别中，维持时空和结构完整性至关重要。保留骨骼数据动态和顺序特性的增强策略对于在基于手势的应用中实现可靠的识别性能至关重要。本文提出了一个专为基于骨架的手势识别而设计的数据增强框架。该框架结合了几何变换、随机裁剪和基于强度的修改，以应对数据集的多样性和现实世界的变异性。通过使用基准数据集 DHG14/28、SHREC17 和 JHMDB 评估增广的数据集，并使用先进的模型和两个竞争力较强的模型——e2eET、FPPR-PCD 和 DD-Net [? ? ?]。通过解决数据增强与基于骨架的手势识别之间的互动，本研究填补了该领域的重要空白，为可扩展、健壮并适用于真实世界的手势识别解决方案铺平了道路。

过去，手势识别方法传统上依赖于 RGB 图像或深度数据来捕捉动作，卷积神经网络 (CNN) 提取时空特征。虽然 2D CNN 用于静态手势，但 3D CNN 结合时间维度用于动态手势 [? ?]。然而，基于图像的技术面临着光照变化、遮挡和背景混乱等挑战。为了解决这些问题，已经开发了集成深度、骨骼数据和点云的多模态方法，通过结合空间和深度信息提高了鲁棒性 [? ? ?]。

e2eET 骨架基于数据级融合的手势识别 [?] 使用不同数据流的学习来表示同一手势的不同拓扑结构，例如自上而下、侧左或前后方向。通过以端到端的方式微调 CNN 流的集合，该系统在语义表示和数据集成方面实现了更深入的理解，从而实现更精确和更稳健的手势检测和识别。e2eET 模型目前在 SHREC'17 [?] 和 DHG 数据集 [?] 上是最先进的 (SOTA) 模型。[?] 提出了基于双流框架的点云手势识别，该框架整合了原始点云和残差基点集 (BPS)，以解决手势识别中的多尺度挑战。通过以 DenseNet 和点网络组合的方式融合局部和全局的空间手部运动表示，他们的框架在 DHG14/28 和 SHREC' 17 [? ?] 等数据集上的表现优于传统方法。FPPR-PCD 模型在 SHREC'17 数据集上的性能排名第二。[?] 将手势识别作为不规则序列学习任务，并提出了一种 PointLSTM 来对点云序列中的动态运动进行建模，并通过邻域分组有效传播来自之前帧的信息，以保持空间一致性。该方法在 SHREC' 17 和 NVGesture [?] 数据集上进行了评估，并展示

了其有效利用运动和形状特征的能力。通过克服传统图卷积网络（GCN）的缺陷，时间解耦图卷积网络 [?] 引入了时间相关的邻接矩阵，允许时间敏感的拓扑学习。这种方法与通道依赖邻接矩阵相结合，提供了对手势序列中时空关系的理解。

DD-Net 是一个用于基于骨架的动作识别的轻量且高效的架构。DD-Net 解决了之前方法中观察到的大模型尺寸和执行速度慢的问题。通过制定新策略，如双特征双运动策略，其中位置视角的关节收集集合距离（JCD）也与描绘快慢时间变化的两种尺度的全局运动特征相结合。DD-Net 使用 SHREC'17 和 JHMDB 数据集进行评估。硬件的进步，如边缘计算，以及对抗性训练等新兴范式，进一步推动了在游戏、增强现实和医疗辅助技术等应用中的实时手势识别。然而，仍然存在挑战，包括遮挡和用户间的差异性，这些问题正通过元学习和领域适应等技术来解决。为了提高复杂神经网络的泛化能力，不同的数据增强技术被研究。SimplePairing 提出了一种增强框架，通过平均像素值混合从训练集中随机选择的两个图像来研究其影响，这产生了 N^2 个新训练样本，并在 CIFAR-10 中将 top-1 错误率降低了 1.29 个百分点，在使用 GoogLeNet 的 ILSVRC 2012 数据集中降低了 4.5 个百分点。[?] 引入了一种非常规的数据增强技术，通过裁剪四张不同的图像并将它们拼接在一张图像中来创建新的训练图像，并对裁剪图像类别进行软标记。该方法在 CIFAR-10 上实现了 2.19 % 的测试错误率 [?]。为了增强神经网络的泛化能力，[?] 提出了一种增强技术，即 cutMix，它用另一个图像的补丁替换某图像的区域，同时按比例混合它们的标签，以保留在传统区域丢失导致的信息像素。CutMix 在 ImageNet 分类上的 top-1 准确率有所提升，分别在 ResNet-50 和 ResNet-101 架构上达到 2.28 % 和 1.70 % 的提升 [?]。

数据增强旨在模拟现实场景并提高 CNN 的泛化能力。[?] 中提出的增强方法包括通过用随机像素值擦除图像的矩形部分以模拟现实环境中常见的遮挡。这种方法在 WRN-28-10 网络架构 [?] 上展示了效果，使 Fashion-MNIST 的 top-1 错误率减少了 0.36 %，并且使用 ResNet-110 网络架构 [?] 的精确度提高了 0.49 %。

最近，人们开发了复杂的增强技术，利用自动化和专业知识来生成多样且逼真的数据。Mixup [?] 通过在图像及其标签之间进行线性插值生成增强样本，从而提高模型的泛化能力。此外，CutOut [?] 提出了一种方法，通过遮盖图像的随机矩形部分，使模型专注于全局上下文而非局部特征。

2. 实施

提出的框架通过实现一个数据增强流程来增强手势识别，该流程解决了诸如数据集大小有限、视角偏差和环境变化等挑战，同时保留手势特定特征。图 1 是 AugmentGest 框架的流程。

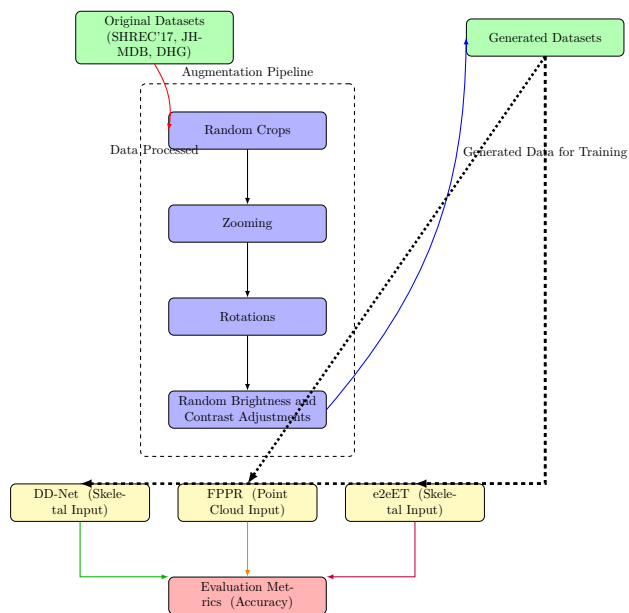


Figure 1. 框架概述：所提出的增强框架提升了训练数据和模型在不同数据集与架构上的性能。

2.1. AugmentGest 框架

设 $x \in \mathbb{R}^{W \times H \times C}$ 和 y 分别表示一个训练图像及其标签。提出的增强算法的目标是通过将单个输入样本 (x, y) 应用多种变换来生成一个新的训练样本 $(\tilde{x}, \tilde{y})$ 。生成的训练样本 $(\tilde{x}, \tilde{y})$ 用于以其原始损失函数训练模型。

该算法旨在将训练样本数量增加四倍，从而显著增强数据集的多样性。施加于数据的增强序列描述如下。序列的转换可以表示为：其中， T_{crop} 、 T_{rotation} 、 T_{zoom} 和 $T_{\text{brightness, contrast}}$ 分别代表裁剪、旋转、缩放和亮度/对比度变换。图 ?? 显示了对 SHREC'17 数据集中的“抓取”手势样本应用所提出框架的结果。通过数据级融合，深度图像和 22 个关节坐标合并成表示手部关节位置的静态二维时空图像。时序信息通过颜色强度编码，如 [?] 所述。与诸如 CutMix [?]、MixUp [?] 和 SimplePairing [?] 之类的混合训练样本的增强方法不同，我们的框架保留了手势特定的空间和时间完整性。随机裁剪确保保留关键区域，而旋转、缩放和亮度/对比度调整提高了对方向、比例和光照变化的鲁棒性。通过保持结构一致性并使训练数据集增加四倍，我们的方法提供了更丰富的样本集以进行鲁棒的模型训练。表 ?? 比较了在 DD-Net 模型上的 MixUp 和 CutMix。尽管基线实现了 81.82 % 的准确率，MixUp 和 CutMix 分别导致 71.50 % 和 72.50 %，分别显示出 10.80 % 和 9.09 % 的性能下降。这些方法虽然在其他领域有效，却破坏对于手势识别至关重要的时空一致性。

2.2. 数据集

AugmentGest 在三个基准数据集上进行手势和人体动作识别的评估。SHREC'17 [?] 包含 2800 个手势序列, 跨越 14 和 28 个手势类别, 而 DHG14/28 [?] 包含 20 个受试者执行的 2800 个序列, 涉及 14 个手势类别。JHMDB 数据集 [?] 包含 928 个动作视频, 涵盖 21 个人体动作类别, 并附有关节数据注释。这些数据集共同提供了一个多样化的环境来验证所提出的增强策略。

2.3. 实验装置

实验在使用云资源的 Kaggle Notebooks 上进行。DD-Net 和 e2eET 在没有 GPU 加速的情况下运行, 展示了其在低功耗环境中的适用性, 而 FPPR-PCD 由于其计算复杂性需要一个 GPU。使用两个 NVIDIA Tesla T4 GPU (T4 x2) 进行 CUDA 加速, 展示了在不同计算环境下的可扩展性和效率。表 ?? 比较了再现的基线性能与所提出的增强框架的结果。已发表的 SOTA 结果略高, 然而, 再现结果与原始实现一致, 提供了一个可靠的基线来评估框架的影响。

2.4. 手势识别框架的评估

所提出的框架将增强数据集整合到各种手势和运动识别模型中, 以将动态手势序列 g_i 分配至手势类别 C_h 中的正确类别 $S = \{C_h\}_{h=1}^N$ 。评估的模型包括:

- E2eET [?]: 一种基于多流 CNN 的架构, 使用来自多个视角的手势表示的数据级融合。评估中重现了使用 Adam 优化器和同方差交叉熵损失的 PyTorch 实现, 利用了 ResNet34。准确性监控包括自上而下、定制和面朝前视角。
- DD-Net [?]: 一种轻量级模型, 专为从 3D 骨骼关节数据中学习空间和时间特征而优化, 采用分类交叉熵损失和 Adam 优化器进行训练。
- FPPR-PCD [?]: 一种双流点云模型, 能够解耦局部姿态和全局空间特征。通过 dbscanCluster 格式将深度图转换为点云序列。

此设置反映了一个精心优化的流水线, 旨在提高各种架构的性能, 验证其在提高手势识别准确性方面的多功能性和有效性。

2.5. 结果与讨论

所提出的框架通过引入尺度、方向的变化来扩展数据集的大小和多样性, 同时保留关键的手势特征。对于 SHREC'17 14G 数据集, 当使用 AugmentGest 进行扩充数据样本训练时, 最先进的 e2eET 模型表现出显著的性能提升。如表 ?? 所示, e2eET 模型实现了 1.54 % 的准确率提升, 超越了之前报道的在 [?] 的 SOTA 结果。该框架在 SHREC'17 28-gesture 和 DHG 28-gesture 数据集上分别提供了 0.47 % 和 1.31 % 的提升。此外, FPPR 模型受益于 AugmentGest 框架, 使得在 SHREC'17 14 手势数据集上的准确性提高了 0.5 %。这一结果强调了该框架对多样化手势识别框架架构的适应性。表 ?? 中的结果展示了所有模型和数

据集的一致性能改进。值得注意的是, DD-Net 模型在 JHMDB 运动数据集上的准确性提升了 4.54 %, 并且在 SHREC'17 14 手势数据集上的准确性从 94.76 % 提高到了 95.48 %。这些发现验证了所提增广框架在提升不同数据集和架构的手势识别性能方面的有效性, 彰显了其多样性和在各种应用中的潜力。图 2 展示了在 SHREC'17 14G 数据集上训练的模型相对增强数据集的验证准确性。通过 AugmentGest 框架生成的数据集比原始数据集大四倍, 从而允许对模型性能进行更为强健的评估。对于评估, 模型在增强数据集上训练了 200 个周期, 而在原始数据集上训练 600 个周期, 反映了由于训练样本的多样性和数量的增加而减少了延长训练的需求。结果表明, 增强数据集在更少的周期内实现了始终更高的验证准确性, 强调了增广框架在提高模型泛化和效率方面的有效性。此外, 在增强数据集上进行 600 个周期的延长训练将准确性提升到 86 %。

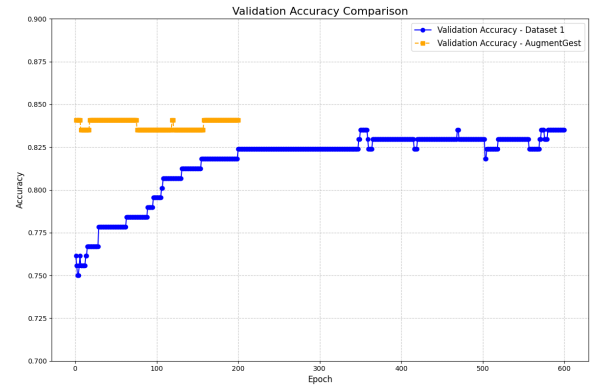


Figure 2. 在不同数据集上训练的 DD-Net 验证准确性: JHMDB 数据集 (数据集 1) 和增强数据集 (AugmentGest)。

2.6. 消融研究

Table 1. 对增强管道的消融研究, 以显示每个步骤对 E2eET 多流 CNN 模型在 SHREC'17 上的性能的贡献。

Framework Step	Accuracy (%)
No Augmentation (Baseline)	96.67
Image Crops	97.50
Image Rotations	97.50
Image Zoom	96.55
Brightness & Contrast Adjustment	97.14

为了进一步研究所提出框架的影响, 使用 SHREC'17 数据集对 E2eET 模型 [?] 的个体增强步骤进行了评估, 如表 1 所示。图像的裁剪和旋转分别提高了 0.83 %, 亮度 & 对比度调整提高了 0.47 %, 而图像缩放没有影响。结合所有步骤之后, 精确度达到了 98.20 %, 突显了增强技术的互补益处。

Table 2. E2eET 模型的运行时。

Dataset	Original	Augmented
SHREC'17 14g	03h:04m:26s	03h:10m:42s
SHREC'17 28g	02h:53m:45s	03h:09m:07s
DHG 28g	02h:54m:37s	02h:52m:53s

2.7. 训练时间评估

表 2 展示了与使用 AugmentGest 框架生成的数据集相比，基于 2eET 骨架的 HGR 在数据集上的训练和评估运行时间。尽管增强后的数据集大了四倍，运行时间的增加却很小。对于 SHREC'17 数据集 14G 和 28G，运行时间仅增加了几分钟，而对于 DHG 28 手势数据集，运行时间几乎保持不变。这些结果证明了所提出增强框架的计算效率，使其成为一个合理的解决方案，可以显著扩展数据集规模而不会带来大量开销。

本文介绍了一种新颖的增强框架，旨在提高手势识别模型的性能。通过保持时空连贯性和骨架特征，该框架在不增加数据收集成本的情况下提高了准确性。通过整合多样的增强策略，该框架在 e2eET、DD-Net 和 FPPR-PCD 等最先进的模型上，在 SHREC'17 和 DHG 等基准数据集上展示了显著的准确性提升。这些结果强调了该框架即便在复杂性和手势多样性不同的数据集上，也能增强模型泛化和效率的能力。虽然框架展示了其潜力，但仍然存在挑战，特别是在应对现实世界的变化和扩展其在更多样化数据集上的应用方面。未来的研究可以集中于上下文感知的增强技术和注重注意力的方法，以进一步完善和优化该框架。通过解决这些挑战，这种方法可以为更强大和适应性强的手势识别系统在动态和复杂的环境中铺平道路。本研究的发现强调了数据增强在推进手势识别中的关键作用，帮助提高了包括游戏、虚拟现实和辅助技术在内的多个应用领域的鲁棒性和准确性。