基于 UConvLSTMs 网络的批量运动目标捕获焦距预测方法

李 超[1] 张 攀[1]
LI Chao ZHANG Pan

摘 要 针对现有以位置检测为主的目标检测方案在采集批量运动目标时，存在大量细节缺失问题，文章提出了一种可支持光学变焦倍数预测的网络模型，辅助形成细节捕获为主的目标检测方案。该模型以 UNet 网络与 ConvLSTM 网络为核心而改进，通过与变焦倍数关联的图像分割数据标签的监督训练，实现利用 n 帧历史图像直接预测第 n+1 帧图像中附属主体的前景分割区域，从而间接获得批量目标细节采集所需变焦倍数。仿真实验表明，网络模型针对人脸和 QR 码的焦距预测误差容限为 5% 时准确率分别达到 73%、61%。

关键词 焦距预测；细节捕获；ConvLSTM；UNet

doi：10.3969/j.issn.1672-9528.2025.02.020

0 引言

目标检测是图像处理领域最热门的研究内容之一，尽管现有的模型算法丰富，但其主要适用于数字变焦场景，且基于这些算法所构建的方案，仍然以位置检测为主、细节捕获为辅，在工程应用时会面临诸多挑战，尤其是针对批量多尺度运动目标细节进行实时有效检测与采集时的效果可能不佳。如在公共场所流动人群人脸采集场景中，不仅需要快速完成多种尺寸的批量人脸检测与定位，还要求其根据检测定位结果采集到的人脸细节信息，用以支撑后续失信人员追踪、可疑人员匹配等特殊情况的执行；在批量运动 QR 码细节采集场景中，不仅需要完成较小尺寸的批量 QR 码的检测与定位，还要求其根据检测定位结果采集到的目标细节信息，支撑后续解码业务的执行。

对于输出高分辨率图像的数字变焦系统的研究，在现有算法构建的先前级检测目标再后级获取细节的方案中，还存在部分尺度目标漏检、流程效率低下问题。受相机物理参数、目标有效性所需最低分辨率等限制，大尺度图像中可获得有效细节的多尺度目标数量存在最优值。同时，为了提高多尺度目标检测准确率，多种特征融合组件被添加到方案前级步骤的目标检测模型中 [1]。然而后级步骤中仍然需要基于前级检测定位结果进行反复采样，并根据采样后所获得的细节信息有效程度进行目标有效区域的反复定位。这种反复采样与定位过程可能加重系统执行效率低下问题，并导致系统单次批量多尺度目标检测召回率降低。

对于输出中等分辨率图像的光学变焦系统，由于其变焦过程仍然以机械结构为基础而实现，因此不适宜执行数字变焦中常用的反复采样过程。相关研究主要基于单帧历史图像进行变焦倍数预测，效率高，但性能往往不足，更适合直接嵌入光学变焦摄像头内部，执行相关业务的预处理工作 [2]。近年来随着人工智能技术的发展，基于深度学习的变焦预测方法逐步成为主流，使得在被拍摄目标持续运动时的目标细节采集效果持续改善 [3-4]。

因此，本文以细节捕获为主、位置检测为辅，基于分割与时序网络，提出一种变焦预测模型，以辅助批量目标细节采集。即在应用方案的前级步骤中，利用深度卷积神经网络模型，直接预测相机捕获含有合适目标细节图像所需焦距信息，从而在多个连续运动目标采集场景中，间接实现后级步骤中检测定位与细节捕获的同步。

1 本文方法

1.1 变焦预测模型 UConvLSTMs

变焦预测模型 UConvLSTMs 由 2n 个相同的 UConvLSTM 模块组合而成，如图 1 所示，其中 n 个模块构成编码器，以提取记忆输入 n 帧图像的有效信息，另外 n 个模块构成解码器，并分别预测输出后续时刻目标对象的变焦偏移量。UConvLSTM 模块由 U-Net[5] 与 ConvLSTM[6] 改进而来。其中 U-Net 网络输入尺寸为 512 px×512 px，共进行四级尺寸缩放过程。S 操作参考空间注意力机制实现，用公式表示为：

$$S=\mathrm{softmax}(w_{FS}*[\mathrm{MaxPool}(F_{chnal\text{-}refined}),\mathrm{AvgPool}(F_{chnal\text{-}refined})]) \quad (1)$$

首先对特征图在通道维度进行平均池化和最大池化后进行拼接，然后再利用 softmax 函数进行压缩。T 操作则为 tanh 函数，是将输入特征归一化到 [-1,1] 区间。$F_{chnal\text{-}refined}$ 表示按照通道数进行处理后的输入特征图；w_{FS} 表示全连接运算。

1. 内江师范学院人工智能学院 四川内江 641100
[基金项目] 内江市东兴区经济和信息化局科研项目（QKJ202103）

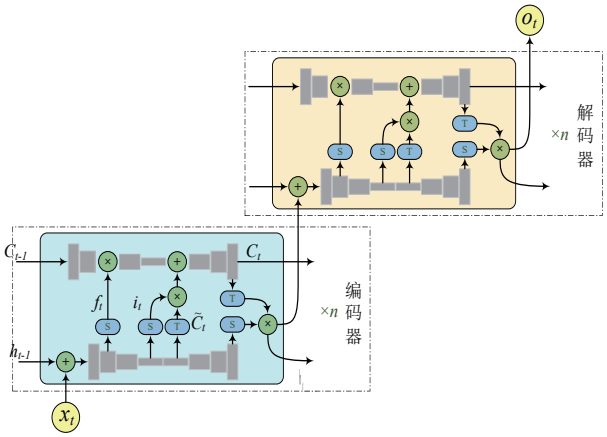

图 1 UConvLSTMs 模型基础结构

UConvLSTM 模块的主要计算过程为：

$$\begin{aligned} f_t &= S(W_{\text{xhf}} * (x_t + h_{t-1}) + b_f) \\ i_t &= S(W_{\text{xhi}} * (x_t + h_{t-1}) + b_i) \\ \tilde{C}_t &= T(W_{\text{xhc}} * (x_t + h_{t-1}) + b_c) \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \\ h_t &= S(W_{\text{xho}} * (x_t + h_{t-1}) + b_o) \circ T(C_t) \end{aligned} \tag{2}$$

式中：* 表示卷积运算；∘表示乘法运算。

由于本文设计的数据标签是未来第 n+1 帧图像中的目标轮廓信息，因此，输入特征 x_t 中包含的是基础变焦特征，隐藏特征 h_{t-1} 中包含的是偏移变焦特征。因此，遗忘门 f_t 是对输入基础变焦特征 x_t 和偏移变焦特征 h_{t-1} 相加后进行的 S 操作；更新门的 i_t 和 $\tilde{C}_t$ 分别是对输入基础变焦特征 x_t 和偏移变焦特征 h_{t-1} 相加后进行的 S 操作与 T 操作；遗忘门与更新门作用后，记忆信息的输出为 C_t；模块的输出 h_t 则是当前输入图像信息与历史记忆信息共同作用后的结果。W_{xhf}、W_{xhi}、W_{xhc}、W_{xho} 分别代表相应输入输出训练后的网络参数，b_f、b_i、b_c、b_o 分别代表与前面训练参数对应的偏移量。

1.2 模型预测变焦倍数原理

为了更有效采集批量目标，如图 2（a）所示，人脸检测附属主体设定为行人躯干，如图 2（b）所示，QR 码检测附属主体设定为手持终端，则采集系统可抽象为图 2（c）所示。采集系统获得图像画面中，可采集细节信息的数据标签包括：图像中最小面积目标的缩放倍数 ι_a、最右侧目标有效检时的缩放倍数 ι_b、最左侧目标有效检时的缩放倍数 ι_c 等。

（a）人脸采集环境（b）QR 码采集环境（c）采集环境俯视图

图 2 采集环境及采集样例

根据实验环境，UConvLSTMs 模型构成系统进行焦距预测的原理可抽象如图 3 所示。本质上即利用前 n 帧的图像输入 UConvLSTMs 模型预测输出第 n+1 帧图像中附属主体的面积大小，从而确定在第 n+1 帧画面中应该缩放的倍数，使得在该画面中获得有效目标最多。

图 3 系统预测焦距原理

1.3 数据标签生成方法

UConvLSTMs 所需数据标签是对传统图像分割数据标签的改进，其生成过程主要包括三个步骤：

1.3.1 确定画面中有效目标的最小面积和最大面积

根据采集场景的图像画面，将可进行后续处理所需目标的最小面积确定为常量 B 和最大面积确定为常量 M。

1.3.2 标记可采集细节信息的所有目标

所有目标变焦缩放下限倍数 ι_α 按公式计算为：

$$\iota_\alpha = \varepsilon(t) \cdot \frac{B}{t} \cdot k \tag{3}$$

式中：缩放下限倍数 ι_α 表示小尺寸目标的最小放大倍数，或者大尺寸目标的最大缩小倍数；$\varepsilon(t)$ 为阶跃函数；t 表示目标面积。例如，当目标 t 的面积为 0.5B 时，计算放大倍数 ι_α=2k，当目标 t 的面积为 2B 时，计算缩小倍数 ι_β=0.5k，即当系数大于 1 时表示放大，当系数小于等于 1 时表示缩小，由于是同一计算方法获得，所以所有缩放下限倍数可直接进行比较。

所有目标变焦缩放上限倍数 ι_β 按公式计算为：

$$\iota_\beta = \varepsilon(M - t) \cdot \frac{M}{t} \cdot k \tag{4}$$

式中：缩放上限倍数 ι_β 表示小尺寸目标的最大放大倍数，或者大尺寸目标的最大放大倍数，$\varepsilon(M-t)$ 为阶跃函数，t 同样表示目标面积。例如，假定 M=4B，当目标 t 的面积为 0.5 计算最大放大倍数 ι_β=8kB 时，计算最大放大倍数 ι_β=8k，当目标 t 的面积为 2B 时，计算最大放大倍数 ι_β=2k，同样可知系数大于 1 时都表示放大，且都是同一计算方法，计算结果可以相互比较。

对于缩放倍数计算，若该目标面积小于 B，则是按照式 3 计算的缩放下限倍数，否则是按照式（4）计算的缩放上限倍数。

由于相机物理参数限制的原因，采集图片中可能客观存在后续无法处理的目标，因此数据标注时需要剔除这些目标，从而提高单次检测定位出有效目标的效率，为此设计了可采集细节信息的目标的最优数量计算方法，该算法流程如图 4 所示。

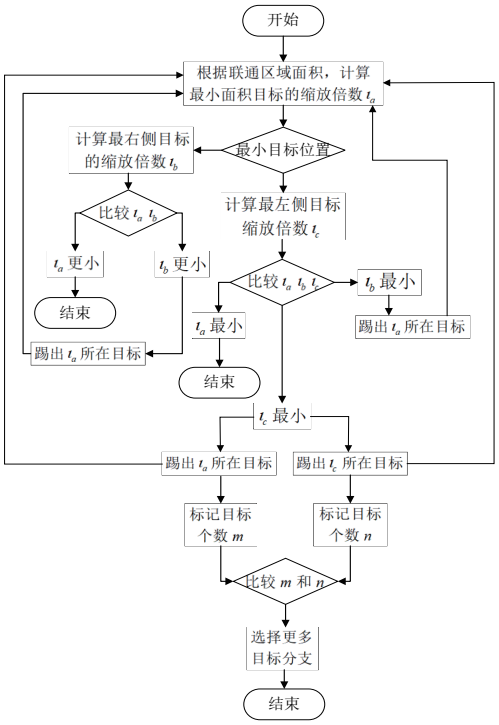

图 4 可采集细节信息的目标的最优数量计算方法

算法将根据联通区域面积，先搜寻到最小面积目标的缩放倍数 ι_a。若最小目标在画面最左侧，则比较最小目标缩放倍数 ι_a 和画面最右侧目标的缩放倍数 ι_b，比较 ι_a 与 ι_b，获得更小值。若最小值为 ι_a 则表明此时所有目标都可以直接进行 ι_a 倍数的缩放，若为 ι_b，则标记 ι_a 所在目标后踢出，重新执行算法。若最小目标不在画面最左侧，则计算画面最左侧目标缩放倍数 ι_c，并对 ι_a、ι_b、ι_c 由大到小进行排序。若 ι_a 最小，即排序是 $\iota_b\iota_c\iota_a$ 或 $\iota_c\iota_b\iota_a$，则表明此时所有目标都可以直接进行 ι_a 倍数的缩放。若 ι_b 最小，则标记 ι_a 所在目标后踢出，回到算法起始点继续执行。若 ι_c 最小，或者则标记 ι_a 所在目标并踢出后，回到算法起始点继续执行，并记录最终标记目标个数 m；或者标记 ι_c 所在目标自身并踢出后，回到算法起点继续执行，并记录最终标记目标个数 n。比较 m 和 n，选择更多标记目标分支路径相关目标即为需要标记的所有可采集细节信息的目标。

1.3.3 生成有效目标附属主体的数据标签

参考图像实例分割数据标记方法，将选定的所有有效目标对应的附属主体区域标记为 1，重合部分进行叠加，背景标记为 0。

按照式（3）（4）计算所有有效目标的缩放倍数范围，及计算获得缩放下限倍数与缩放上限倍数构成的区间，即面积小于 kB 的目标计算获得最小放大倍数与最大放大倍数构成的区间，面积大于 kB 的目标计算获得最大缩小倍数与最大放大倍数构成的区间，其中 k 表示附属主体目标相对真实目标的缩放倍数。按照图 5 所示，首先将当前相机的变焦倍数等价为 1 倍变焦倍数，进而可获得当前相机等价变焦缩放范围为 (0, max]；随后计算所有目标缩放倍数范围区间的交集，并获得交集的中心值，该中心值即为当前模型预测的最终变焦倍数。

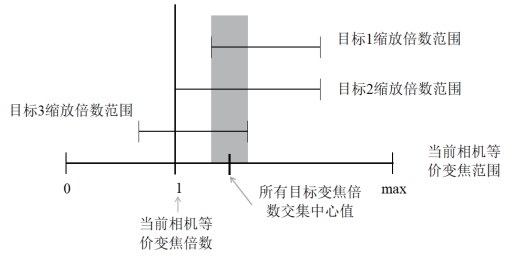

图 5 变焦倍数计算原理

2 数据集与评价指标

2.1 数据集构建及训练环境

按照图 2（c）所示搭建采集环境，模拟真实部署中的光学变焦相机采集的不同焦距下的可见光图像。每组数据采集时，将可采集目标细节的图像所在时刻标记为预测帧图像，并以其为终点，以相同时间为间隔，在录制的视频中抽取 n 帧图像，最终两种场景分别收集 3500 组数据，且标注按照 1.1 节中的方法进行。

模型训练测试硬件环境：Tesla V100 GPU、Intel Xeon CPU、64 GB 内存硬件环境，软件环境为 Windows Server 操作系统、PyTorch 1.5 深度学习框架。数据集按照 6:2:2 区分为训练集、验证集、测试集。

2.2 实验评价指标

实验对模型输出相关的两类对象进行评价。第一类评价对象针对目标预测像素区域，主要参考图像分割评价标准，指标 TPR、TNR 计算公式分别为：

$$\mathrm{TRP}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{6}$$

式中：TP 表示正确预测为正例的样本数；TN 表示正确预测为负例的样本数；FN 表示错误地将正例预测为负例的样本数；FP 表示错误地将负例预测为正例的样本数。

第二类评价对象针对预测的焦距信息。尽管训练数据的标签和网络模型的预测输出焦距都是单一的确定值，但实际使用预测焦距缩放目标图像时，焦距在一定范围内都是有效的。为了更全面的评价网络模型，按照式（7），进行 3 个分段区间的预测准确率计算。

$$P_{\mathrm{interval}}=\frac{R}{T} \tag{7}$$

式中：R 表示实验预测正确次数；T 表示总的实验次数；$P_{5\%}$、$P_{10\%}$、$P_{20\%}$ 分别表示预测焦距值在 5%、10%、20% 以内认定为正确时相应的准确率。

3 实验与讨论

3.1 基于目标像素区域预测的序列帧数选择实验

为了对 UConvLSTMs 模型中编码器与解码器的 UConvLSTM 模块数量进行设计，也就是确定最佳的输入序列帧数，本文进行了从 10 帧到 1 帧的实验。也就是从视频图像中等间隔的抽取 n+1 帧图像，并将前 n 帧作为输入序列，第 n+1 帧图像作为预测目标。

对于批量人脸细节采集场景，获得的目标像素区域的预测指标 TPR、TNR 的量化结果如图 6 所示。

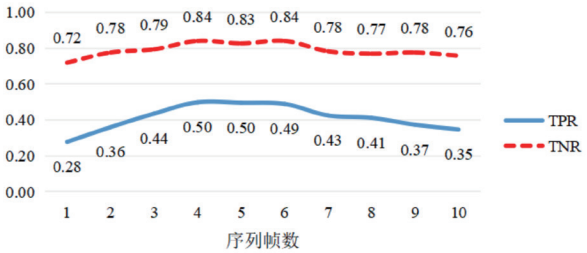

图 6 人脸细节采集场景实验量化结果

结合分析可知，当输入序列图像低于 3 帧时，模型性能随帧数明显提升，而当输入图像序列高于 6 帧时，模型预测质量会出现一定的下降。结合采集场景中人群移动速度、实验平台性能等分析可知，较少序列帧会使网络模型提取运动特征困难，较多序列帧，即更密集的采样图像，要求实验平台具有更快地响应速度，且往往快于光学变焦过程。同时，尽管选择 5 帧和 6 帧图像作为输入时模型的预测性能也较高，但相对 4 帧输入提升并不明显，考虑到降低模型参数量的需求，因此该场景下 UConvLSTMs 网络模型最佳的模块数量选定为 4，即编码器与解码器都采用 4 个 UConvLSTM 模块级联。

对于批量 QR 码细节采集场景，获得的目标像素区域的预测指标 TPR、TNR 的量化结果如图 7 所示。由于 QR 码采集的附属主体目标为手持终端，其移动时相对身躯的变化频率更高，结合试验结果，选定 5 个 UConvLSTM 模块级联分别构成 UConvLSTMs 模型的编解码器。

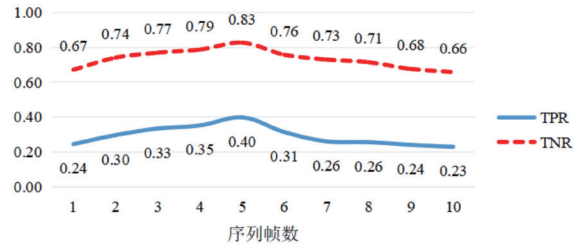

图 7 QR 细节采集场景实验量化结果

3.2 焦距预测实验

焦距数值预测实验结果如图 8 所示，可以看出针对人脸细节识别所需焦距预测相对 QR 码细节识别所需焦距预测更准确，这是由于 QR 码本身相对人脸较小，而其解码所需分辨率却不低。

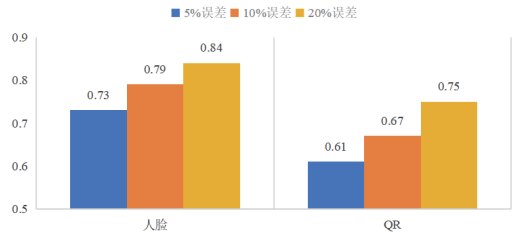

图 8 焦距预测实验量化结果

4 结论

针对目标检测中的细节捕获问题，本文基于 UConvLSTMs 网络提出了一种批量运动目标捕获焦距预测方法。方法核心网络由多个以 UNet 改进的 LSTM 构成，在获得目标分割信息的同时，对序列图像进行目标捕获所需焦距的预测。实验结果证明了方法的有效性，但在目标尺寸较小时，其焦距数值预测的绝对准确率仍然较低，因此后续研究中将着重提升小目标细节捕获所需焦距预测性能。

参考文献：

[1] 王利祥，郭向伟，卢明星 . FPN 算法在视觉感知机器人抓取控制的应用研究 [J]. 机械设计与制造，2024(4): 303-307.

[2] 杨小君，苏秀琴，郝伟，等 . 用于精密测量的自动变焦系统及标校方法的研究 [J]. 光子学报 ,2005(12):1921-1924.

[3] 赫贵然，李奇，冯华君，等 . 基于 CNN 特征提取的双焦相机连续数字变焦 [J]. 浙江大学学报 (工学版),2019,53(6): 1182-1189.

[4] 宋炯辉，李奇，王静，等 . 基于纹理修复的双焦相机连续数字变焦算法 [J]. 浙江大学学报 (工学版), 2021,55(8):1510-1517.

[5] 宋廷强，刘童心，宗达，等 . 改进 U-Net 网络的遥感影像道路提取方法研究 [J]. 计算机工程与应用 ,2021,57(14):209-216.

[6] 周玮辰，韩震，张雪薇 . 基于融合 U-Net 及 ConvLSTM 的海面高度异常预报方法研究 [J]. 海洋通报 ,2021,40(4):410-416.

【作者简介】

李超（1991—），女，四川内江人，硕士，讲师，研究方向：交叉学科、优化技术。

张攀（1989—），通信作者（email:2358792637@qq.com），男，四川资阳人，硕士，讲师，研究方向：交叉学科、图像处理。

（收稿日期：2024-10-26）