[文章编号] 1003－4684(2020)02-0052-04

一种基于无监督学习的空间域图像融合方法

王淑青[1,2]，蔡颖婧[1,2]

(1 湖北工业大学湖北省电网智能控制与装备工程技术研究中心，湖北 武汉 430068；
2 湖北工业大学电气与电子工程学院，湖北 武汉 430068)

[摘 要] 针对多焦点图像融合问题，提出一种新的无监督深度学习模型。首先，训练一个无监督的编解码器网络来提取输入图像的深层特征，然后利用这些特征和空间频率来测量图像像素活跃度并得到决策图。最后，应用一致性验证方法对决策图进行调整，得出融合结果。该方法的关键在于，只有在景深(DOF)范围内的物体在照片中才有明显的锐度，而其他物体很可能是模糊的。本方法是在深度特征上分析锐度的特征，而不是原始图像。实验结果表明，与已有的 16 种融合方法相比，该方法在客观评价和主观评价方面均取得了较好的融合效果。

[关键词] 多焦点图像融合；像素活跃度；深度特征；决策图

[中图分类号] TP391.4 [文献标识码] A

过去的几十年里出现的多种图像融合方法可以分为两类：变换域方法和空间域方法[1]。最经典的是基于多尺度变换的变换域融合方法(MST)理论，如基于拉普拉斯算子的金字塔(LP)[2]和低通的比率金字塔(RP)[3]以及离散小波变换(DWT)[4]，还有双数复小波变换(DTCWT)[5]、曲波变换(CVT)和非下采样轮廓波变换(NSCT)，以及稀疏表示(SR)和抠图融合(IMF)[6-8]。这些方法的关键在于，可以在选定的变换域中通过分解系数来测量源图像的活跃度。显然，变换域的选择在这些方法中起着至关重要的作用。

近年来，深卷积神经网络(CNN)在图像处理方面取得了巨大的成功。一些研究试图使用大容量深卷积模型来测量活跃度。Liu 等[9]首次将卷积神经网络应用于多聚焦图像融合。Prabhakar[10]提出了一种基于 CNN 的无监督曝光融合方法，称为深度融合。Li 和 Wu[11]提出了 DenseFuse 来融合红外图像和可见光图像，采用无监督的编解码器策略来获取有用的特征，并按 L1-norm 融合。受深度融合的启发，本文以无监督编解码器的方式训练网络，并且将空间频率作为融合规则来获得源图像的活跃度和决策图，这与关键假设一致，即只有在景深范围内的对象才具有清晰的外观。

1 图像融合

首先，在训练阶段，训练一个自动编码器网络来提取高维特征。然后利用融合层在融合阶段的深层特征计算出融合层的活跃度。最后，得到了融合两个多焦点源图像的决策图。本文提出的算法只针对融合两幅源图像。

1.1 深层特征提取

从 DenseFuse 得到启发，在训练阶段舍去融合操作，只使用编码器和解码器对输入图像进行重构。在编码器和解码器的参数确定后，利用空间频率从编码器获得的深层特征中计算活跃度。

编码器由两部分(C_1 和 SEDense 块)组成。C_1 是编码器网络中的一个 3×3 卷积层。DC_1、DC_2和 DC_3是 SEDense 块中的 3×3 个卷积层，每层的输出通过级联操作连接。为了精确地重建图像，网络中不存在池层。挤压和激励(SE)块可以通过自适应重新校准 channel-wise 特征响应来增强空间编码，实验表明了这种结构的影响。解码器由 C_2、C_3、C_4和 C_5组成，用于重建输入图像。为了训练编码器和解码器，将损失函数 L 最小化，它结合了像素损失 L_P和结构单线性度(SSIM)损失 L_{SSIM}。其中 λ 是一个常数，体现损失目标的权重

$$L = \lambda L_{ssim} + L_P \tag{1}$$

像素损失 L_P表示输出(O)和输入(I)之间的欧氏距离。

$$L_P = \| O - I \|_2 \tag{2}$$

SSIM 损失 L_{ssim}表示 O 和 I 之间的结构差异，其中 SSIM 表示结构相似操作。

$$L_{ssm} = 1 - \mathrm{SSIM}(O, I) \tag{3}$$

[收稿日期] 2019－09－30

[第一作者] 王淑青(1969－)，女，河北衡水人，理学博士，湖北工业大学教授，研究方向为智能检测与控制，系统分析与集成

1.2　利用深度特征进行空间频率计算

在本文中，编码器为图像中的每个像素提供高维深度特征。但原始的空间频率是在单通道灰度图像上计算的。因此对于深层特征，其修改了空间频率计算方法。设 F 表示由编码器块驱动的深度有限元。$F_{(x,y)}$ 表示一个特征向量，(x,y) 表示这些向量在图像中的坐标。本文使用下面的公式计算它的空间频率，其中 RF 和 CF 分别是行向量频率和列向量频率。

$$RF_{(X,Y)}=\sqrt{\sum_{a=-rb=-r}^{r}\sum^{r}[F_{(x+a,y+b)}-F_{(x+a,y+b-1)}]^2} \tag{4}$$

$$CF_{(x,y)}=\sqrt{\sum_{a=-rb=-r}^{r}\sum^{r}[F_{(x+a,y+b)}-F_{(x+a-1,y+b)}]^2} \tag{5}$$

$$SF_{(x,y)}=\sqrt{\frac{(CF_{(x,y)})^2+(RF_{(x,y)})^2}{(2r+1)^2}} \tag{6}$$

其中 r 为核半径。原始的空间频率是基于块的，而本文方法是基于像素的。于是可以比较两个对应的 SF_1 和 SF_2 的空间频率，其中 SF_k 中的 k 是源图像的索引。

$$D_{(x,y)}=\begin{cases}1,if SF1_{(x,y)}\geqslant SF2_{(x,y)}\\0,otherwise\end{cases} \tag{7}$$

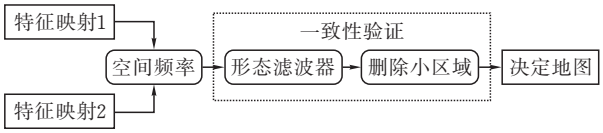


图 1　详细的融合策略

1.3　一致性验证

在连接部分可能有一些小的线或毛刺，并且一些相邻的区域可能被不适当的决定断开。于是使用小型磁盘结构元素交替打开和关闭操作符来处理决策映射。这样既可以消除小的线或毛刺，平滑聚焦区域的连接部分，又将相邻区域合并为一个整体区域。当圆盘结构的半径等于空间频率核半径时，可以很好地检测到小直线或毛刺，并能正确地连接相邻区域。本文采用了与 Liu 等人相同的小区域去除策略并将小于区域阈值的区域反转。本文通常将阈值设置为 $0.01\times H\times W$，其中 H 和 W 分别为源图像的高度和宽度。

在聚焦区域和非聚焦区域之间存在一些不需要的工件。与 Nejati，Samavi 和 Shirani[12] 类似，利用有效的保边滤波、制导滤波提高初始决策图的质量，它可将制导图像的结构信息传递到输入图像的滤波结果中。采用初始融合图像作为指导图像来指导初始决策图的滤波。在这项工作中，实验设置本地窗口半径 4 和正规化参数 ε 为 0.1，引导滤波算法。

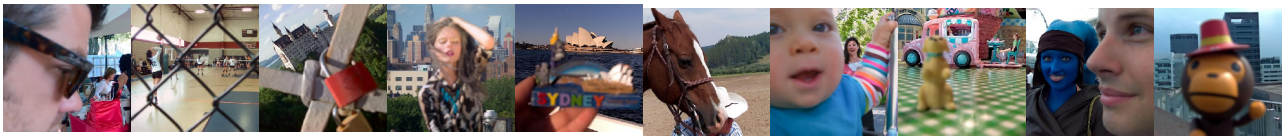
1.4　融合

最后，利用所得到的决策图 D，和像素加权平均规则计算融合后的 F。

$$F_{(x,y)}=D_{(x,y)}\mathrm{Img}1_{(x,y)}+(1-D_{(x,y)})\mathrm{Img}2_{(x,y)} \tag{8}$$

输入图像表示为预先注册的 Imgk，其中 k 表示源图像的索引。融合图像的代表性可视化如图 2 所示。

(a)近聚焦源图像



(b)远聚焦源图像

(c)融合结果

图 2　融合结果的可视化

2　实验

2.1　实验设置

在实验中，使用 38 对多焦点图像作为测试集进行评估。由于无监督策略，首先使用 MS-COCO 训练编解码器网络。该阶段以 82783 幅图像作为训练集，每次迭代使用 40504 幅图像验证重建能力。所有的图像都被调整为 256×256，并转换为灰度图像。学习率设为 1×10^{-4}，每隔 2 个周期下降 0.8 倍。设置 $\lambda=3$ 与 DenseFuse 相同和优化目标函数对权重的网络层。批次大小和年代分别为 48 和 30。然后利用所获得的参数对上述测试集进行 SF

融合。

对该算法的实现源自于公开可用的 Pytorch 框架。网络训练和测试是在一个使用 4 NVIDIA 1080Ti GPU 和 44GB 内存的系统上进行的。

2.2 目的图像融合质量指标

该融合方法是与 16 个代表图像融合方法相比，分别为拉普拉斯算子的金字塔(LP)，低通的比率金字塔(RP)，非抽样轮廓波变换(NSCT)，离散小波变换(DWT)，dual-tree 复小波变换(DTCWT)，稀疏表示(SR)，曲波变换(CVT))，引导过滤(GF)，多尺度加权梯度(MWG)，密集的筛选(DSIFT)，空间频率(SF)的 FocusStack，图像消光融合(IMF)，DeepFuse、DenseFuse (add 和 L1-norm 融合策略)和 CNN-Fuse。

为了客观评价不同方法的融合性能，采用了 Q_g、Q_m 和 Q_{cb} 三个融合质量指标。对于上述三个指标，值越大表示融合性能越好。在本文中可以找到一个很好的全面的质量度量调查。为了进行公平的比较，使用了相关出版物中给出的这些度量的默认参数。

2.3 与其他融合方法进行比较

首先比较了基于视觉感知的不同融合方法的性能。本文主要以两种方式提供了四个例子来说明不同方法之间的差异。

在图 3 中，可视化了两个融合的示例，例如“狮头像”和“笔记本”图像及其融合结果。在每张图像中，聚焦和离焦部分边界附近的区域被放大并显示在左上角。在“狮头像”的结果中，可以看到不同方法对狮头像的边界都进行处理。DWT 显示“锯齿状”形状，CVT、DSIFT、SR、DenseFuse、CNN 显示不需要的工件。对于 DWT 和 DenseFuse，左上角屋檐的亮度也有异常的增加。而 MWG 中相同的区域是失焦的，表明该方法不能很好地检测出聚焦区域。在“笔记本”的结果中，一排挂钟位于聚焦和离焦的边缘，可以看到除了 SESF-Fuse 外，所有的方法都显示出平滑和模糊的结果。

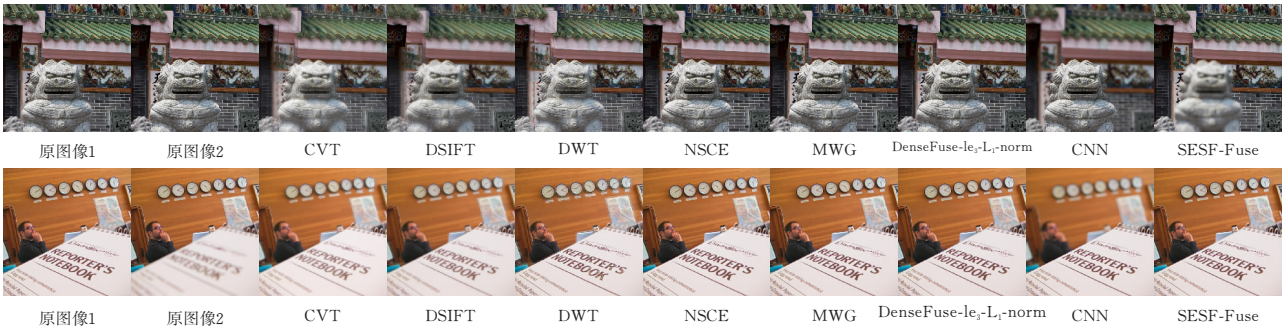


图 3 不同“狮头像”与“笔记本”的可视化融合结果

为了更好的对比，图 4 和图 5 分别显示了从每幅融合图像中减去第一个源图像得到的差分图像，并将每幅差分图像的值归一化为 0 到 1 的范围。如果近聚焦区域被完全检测到，差分图像将不会显示出该区域的任何信息。因此，CVT、DSIFT、DWT 和 DenseFuse-1e3-L1-Norm 不能很好地检测出聚焦区域。SR，MWG 和 CN 在婴儿边缘的区域表现的很好，因为仍然可以看到近聚焦区域的轮廓。此外，SESF-Fuse 在近聚焦区域的中心或边缘区域都有良好的性能。在图 5 中，近焦点区域是石头。与上述观察结果相同，CVT、DSIFT、DWT、NSCT、DenseFuse 不能很好地检测到聚焦区域。除了石头的边界区域，MWG 和 CNN 的效果都很好。

图 4 不同融合方法下“婴儿”的融合结果

图 5 不同融合方法下“海”的融合效果

表 1　与其他融合方法的比较

评价指标	Q_g	Q_m	Q_{cb}
Deepfuse	0.6730(0)	2.4617(0)	0.5650(0)
FocusStack	0.4708(0)	2.8509(0)	0.6330(0)
SF	0.5115(0)	2.8512(0)	0.6023(0)
DenseFuse-1e3-add	0.5189(0)	2.8529(0)	0.6007(0)
DSIFT	0.5266(0)	2.8725(0)	0.6067(0)
DenseFuse-1e3-L1	0.5282(0)	2.8560(0)	0.5972(0)
GF	0.5631(0)	2.8505(0)	0.7007(3)
CVT	0.6186(0)	2.9562(0)	0.6908(0)
DWT	0.6222(0)	2.9465(1)	0.6712(0)
IMF	0.6324 (2)	2.8844(0)	0.7361(4)
RP	0.6478(0)	2.9460(0)	0.7101(0)
DTCWT	0.6529(0)	2.9582(0)	0.7126(0)
NSCT	0.6587(0)	2.9591(0)	0.7168(0)
SR	0.6685(0)	2.9630(2)	0.7334(0)
LP	0.6730(0)	2.9641(8)	0.7352(0)
CNN-Fuse	0.7101(16)	2.9653(7)	0.7839(9)
SESF-Fuse	0.7104(20)	2.8885(14)	0.7848(20)

表 1 列出了使用上述三个指标的不同融合方法的目标性能。可以看到，基于 CNN 的方法和所提出的方法在 Qg 和 Qcb 融合指标的平均得分上明显优于其他 15 种方法。对于 Qg 指标，CN- Fuse 和 SESF-Fuse 的性能相当。然而，CNN-Fuse 是一种监督方法，需要生成不同模糊程度的合成图像来训练一个两类图像分类网络。相比之下，本方法只需要训练一个不需要生成合成图像数据的无监督模型。对于 Qm 度量，SESF-Fuse 的平均核比 LP 小，但是，所提出的方法的第一个数达到了最大值，这意味着它比其他方法具有更强的鲁棒性。

综合考虑以上主观视觉质量与客观评价指标的比较，提出的基于 SESF - Fuse 的融合方法总体上优于其他方法，在多焦点图像融合中表现出了最先进的性能。

3　结论

本文提出了一种无监督深度学习模型来解决多焦点图像融合问题。首先训练一个无监督的编解码器网络来获取输入图像的深层特征，然后利用这些特征和空间频率计算活跃度和决策图进行图像融合。实验结果表明，与现有的融合方法相比，该方法在客观和主观评价方面均取得了较好的融合性能。证明了无监督学习与传统图像处理算法相结合的可行性。另外同样的策略也适用于其他图像融合任务，如多曝光融合、红外融合和医学图像融合。

[参 考 文 献]

[1] Stathaki T. Image fusion: algorithms and applications [C].Elsevier,2011.

[2] Burt P, Adelson E. The laplacian pyramid as a compact image code[J]. IEEE Transactions on Communications,2003, 31(4):532-540.

[3] Stathaki T. Image Fusion: Algorithms and Applications[M]. Academic Press, 2008.

[4] Li H, Manjunath B,Mitra S. Multisensor image fusion using the wavelet transform[J].Graphical Models and Image Processing,1995,57(3):235-245.

[5] Lewis J J, OCallaghan R J, Nikolov S G,et al. Pixel- and region-based image fusion with complex wavelets [J].Information Fusion :Special Issue on Image Fusion: Advances in the State of the Art,2007,8(2): 119-130.

[6] Nencini F, Garzelli A, Baronti S, et al. . Remote sensing image fusion using the curvelet transform[J].Information Fusion: Special Issue on Image Fusion: Advances in the State of the Art,2007,8(2):143-156.

[7] Yang B, Li S. Multifocus image fusion and restoration with sparse representation[C].IEEE Transactions on Instrumentation and Measurement,2010,59(4):884-892.

[8] Li S, Kang X, Hu J,et al. Image matting for fusion of multi-focus images in dynamic scenes[J].Information Fusion,2013, 14(2):147-162.

[9] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network[J]. Information Fusion,2017,36:191-207.

[10] Prabhakar R. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs [C]. In The IEEE International Conference on Computer Vision (ICCV),2017.

[11] Li H, Wu X. Densefuse: A fusion approach to infrared and visible images[C].IEEE Transactions on Image Processing,2019,28(5):2614-2623.

[12] Nejati M, Samavi S, Shirani S. Multi-focus image fusion using dictionary-based sparse representation[J]. Information Fusion,2015,25:72-84.

（下转第 105 页）