3D CNNs 与 LSTMs 在行为识别中的组合及其应用

秦　阳，莫凌飞，郭文科，李　钒
（东南大学 仪器科学与工程学院，江苏 南京　210096）

摘要：基于机器视觉的人体运动识别在视频监控、虚拟现实、医疗护理等诸多领域发挥着重要的作用。结合深度学习中的三维卷积神经网络和长短期记忆神经网络，提出一种融合模型，并与另外两种行为识别模型——长效递归卷积网络和时空域卷积网络，进行了对比，利用公开的 KTH 数据集，进行了实验测试。实验表明，提出的融合模型与长效递归卷积网络和时空域卷积网络相比，对于人体行为图像或视频数据集的学习效果明显，论证了模型的泛化性能和鲁棒性。

关键词：行为识别；深度学习；神经网络；模式识别
中图分类号：TP183　　**文献标识码**：A　　**文章编号**：1000－8829(2017)02－0028－05

Combination of 3D CNNs and LSTMs and Its Application In Activity Recognition

QIN Yang, MO Ling-fei, GUO Wen-ke, LI Fan
(School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: Human activity recognition based on machine vision plays an important role in many fields such as video monitoring, virtual reality, medical care and so on. A new fusion model is proposed by combining the 3D convolutional neural networks and long short term memory neural network in deep learning, and compared with the other two activity recognition models, long-term recurrent convolution network and spatiotemporal convolutional neural network. Test experiments using public KTH data have been done to prove that the combination model has a better learning effect based on human activity image or video data set, which demonstrates the generalization performance and robustness of the model.

Key words: activity recognition; deep learning; neural networks; pattern recognition

人体行为识别在医疗、教育、安全等方面有着重要的研究意义。随着视频记录的推广，基于机器视觉的人类行为识别方法更具有普适性。

传统的行为识别方式集中于选择人体行为的特征量。但是，不同的任务环境下人工选择的特征量具有差异，模型参数不具备泛化性能。近几年，随着深度学习模型的发展，分层抽象的特征提取方式替代了人工选择方式，利用深度学习模型可以消除人工设计特征的盲目性和差异性，实现自动的特征提取工作。卷积神经网络（CNN，convolutional neural networks）是深度学习模型的一种，通过对于输入数据的卷积和子采样，逐渐抽象出上层特征，在静态图像识别领域中取得惊人的识别效果。

Davis 等人认为只凭借单帧的静态图像无法识别人类行为，应将各帧的图像连接成运动图像序列，才能充分提取视频中的行为信息[1]。Elbassuoni S 等人将 CNN 模型拓展到三维，即空间域的二维和时间域的一维，利用三维卷积神经网络，分析帧图像之间时空域关联特征[2]。Wernicke S 建立 3 种不同结构的卷积神经模型对 youtube 视频进行行为识别[3]。实验结果表明，拓展后的模型可以有效识别多帧图像序列。

对于时间序列的分析，深度学习模型中存在较为经典的递归神经网络模型（RNN，recurrent neural network），其特点在于隐藏层的自连接性，输出反馈至下一次输入，是一种动态的时间延迟网络。此外，Sepp Hochreiter 等人为解决 RNNs 存在的梯度消失问题，在传统 RNNs 上引入存储单元改进为长短期记忆递归神经网络（LSTM，long short term memory network），通过

收稿日期：2016－05－08
基金项目：中央高校基本科研业务费专项（2242014R30021）
作者简介：秦阳（1995—），男，江苏南京人，黎族，本科，主要研究方向为深度学习；莫凌飞（1981—），男，湖南邵阳人，博士后，副教授，主要研究方向为多传感器融合和大规模机器学习。

遗忘门控制训练梯度的收敛，保持长期的记忆性[4]。Jeff Donahue 等人将 CNN 和 LSTM 结合，提出长效递归卷积网络（LRCN，long-term recurrent convolution network），并运用在视频识别和视频描述领域[5]。

本文提出一种基于三维卷积神经网络和长短期记忆神经网络的模型。三维卷积神经网络被用来对输入的图像序列做特征提取，获得包含时间信息的特征序列，长短期记忆神经网络对提取后的特征序列进行处理，用 Softmax 层进行分类。

本文利用公开 KTH（Kungliga Tekniska högskolan）数据集对网络结构进行测试。KTH 数据集包含人类行为视频，4 个场景下 6 类动作，25 个实验对象，总共 600 个样本。实验针对 6 类行为进行识别，模型对人体行为视频的识别率到达了 90.7%，这意味着模型可以有效地运用于人类行为识别。

1 模型构架

1.1 总体框架图

视频数据经过采样和预处理之后分成训练数据集和测试数据集，训练数据被用于模型的构建和参数调整，训练完毕后利用测试数据检验模型的性能。系统框图如图 1 所示。

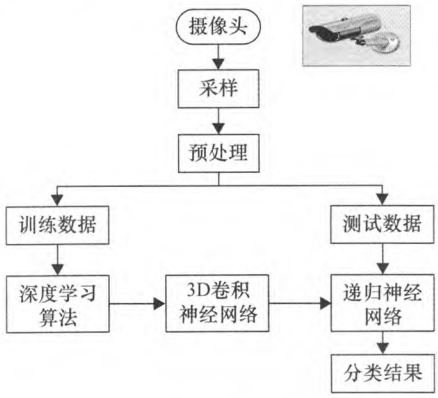

图 1 系统框图

1.2 3D 卷积神经网络

3D 卷积神经网络由卷积层和下采样层交替出现堆叠而成[6-7]，其输入数据是视频中的每一帧按顺序堆叠起来的三维数据[2]。

卷积层中，存在多个卷积核，每个卷积核提取一种特征，卷积核越多生成的特征图越多。第 i 层的第 j 个特征图中位置坐标为 (x,y,z) 的单元的值，由下面公式给出：

$$v_{ij}^{xyz}=f\left(b_{ij}+\sum_{m}\sum_{p=0}^{p_i-1}\sum_{q=0}^{q_i-1}\sum_{r=0}^{R_i-1}\omega_{ijm}^{pqr}v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{1}$$

式中，R_i 为卷积核在时间维度的尺寸；ω_{ijm}^{pqr} 为与第 $(i-1)$ 层第 m 个特征图相连接的卷积核中的坐标为 (p,q,r) 的单元的值。

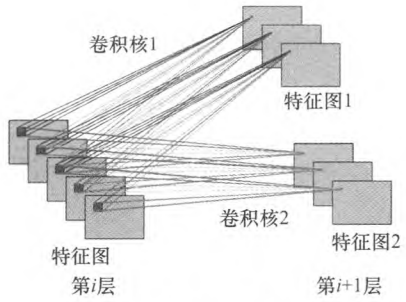

图 2 3D 卷积示意图

下采样层仅仅在空间维度上对特征图进行下采样，合并特征图中临近的特征，特征图时间维度的尺寸不会发生变化。在下采样层中，第 i 层的第 j 个特征图中位置坐标为 (x,y,z) 的单元的值，由下面公式给出：

$$v_{ij}^{xyz}=f\left(b_{ij}+\beta_{ij}down\left(\sum_{p=0}^{n-1}\sum_{q=0}^{n-1}v_{(i-1)j}^{(xn+p)(yn+q)z}\right)\right) \tag{2}$$

式中，z 为时间维度的下标。

1.3 长短期记忆神经网络

递归神经网络是一种隐藏层自连接的深度神经网络，隐藏层的输出不仅到达输出层，还参与下一时间点的隐藏层运算。这种递归结构赋予了模型很深的网络深度。传统的递归神经网络将时间序列的历史信息存储在隐藏层的输出中，隐藏层前向公式可以表示为

$$\begin{cases}h_t=\sigma(W_{xh}x_t+W_{hh}h_{t-1}+b_h)\\ z_t=\sigma(W_{hz}h_t+b_z)\end{cases} \tag{3}$$

网络输出和真实的结果存在残差，为了使输出尽量拟合真实结果，网络通过梯度下降来完成隐藏层中参数值调整。在多层网络中，每层梯度下降可以用链式法则运算。隐藏层梯度下降的公式可以表示为

$$\begin{cases}J=\dfrac{1}{2}(z-z_t)^2\\ W'_{xh}=W_{xh}-\alpha\dfrac{\partial J}{\partial z_t}\dfrac{\partial z_t}{\partial h_t}\dfrac{\partial h_t}{\partial W_{xh}}\\ W'_{hh}=W_{hh}-\alpha\dfrac{\partial J}{\partial z_t}\dfrac{\partial z_t}{\partial h_t}\dfrac{\partial h_t}{\partial W_{hh}}\\ b'_h=b_h-\alpha\dfrac{\partial J}{\partial z_t}\dfrac{\partial z_t}{\partial h_t}\dfrac{\partial h_t}{\partial b_h}\end{cases} \tag{4}$$

式中，z 为真实结果；α 为梯度下降的速率参数，学习速率过大导致学习的不稳定，速率过小导致极长的训练时间，并且可能陷入局部最优解。α 的初始值应设置为 0.8，之后随着残差的缩小不断减少。

传统的递归神经网络存在梯度消失和梯度爆炸两个问题，无法进行长期的训练。长短期记忆递归神经网络通过引入存储单元解决了这两个问题[8]。长短期记忆递归神经网络的隐藏层前向公式可以表示为

$$\begin{cases}i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+b_i)\\ f_t=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+b_f)\\ o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+b_o)\\ g_t=\varphi(W_{xg}x_t+W_{hg}h_{t-1}+b_g)\\ C_t=f_t\cdot C_{t-1}+i_t\cdot g_t\\ h_t=o_i\varphi(C_t)\end{cases} \tag{5}$$

1.4　融合模型

本文中使用的卷积神经网络一共有 4 个隐藏层(hidden layer),整个卷积神经网络中,卷积核大小和池化大小都不变,分别为(3×3×3)和(2×2×2)。文献[2]指出高层特征会随着抽象程度的提高而增加,为了尽可能不丢失特征,3D 卷积网络卷积核数应当逐层增多。池化参数设置为 2,以避免过度池化带来的信息损失。

输入数据为 50 帧的图像序列,对单通道而言,数据规模为 50×60×80,50 为帧数或者序列的长度,60×80 为每个时间戳上图像的大小,由于 RGB 图像存在三通道数据规模需要扩大 3 倍。图 3 为 3D 卷积神经网络各层数据规模。它由 4 层构成,分别是 2 个卷积层和 2 个下采样层。经过 3D 卷积神经网络之后的数据将会被输入递归神经网络进行处理。具体参数的描述如图 3 所示。

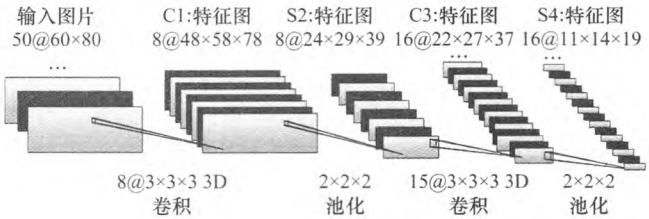

图 3　3D 卷积神经网络及各层参数

C1 层存在 8 个不同 3D 卷积核,接收来自输入的数据,卷积核大小为 3×3×3,特征图的大小为(50-3+1)×(60-3+1)×(80-3+1)=48×58×78,变量的总量为 8×48×58×78=1737216。

S2 层对来自 C1 层的变量做最大池化,池化大小为 2×2×2,选取立方体数据空间 8 个点的最大值作为输出,池化的结果使输入的维度大幅度减少,变量的总量变为 8×24×29×39=217152。

C3 层存在 16 个不同的 3D 卷积核,数据规模变为 16×22×27×37=351648。

S4 层接收来自 C2 的输入,经过池化大小为 2×2×2 的下采样后,最终输出变量规模为 16×11×14×19=46816。

接下来数据将会被调整规模以进入 LSTM 中,3D 卷积神经网络各帧的特征向量填充时间序列,LSTM 按时序中做递归运算,每次递归运算的结果是前面所有特征和当前特征的融合。实际上 LSTM 是在 3D 卷积网络提取的各帧空间信息基础上,获取了帧与帧之间的时间信息。

模型的网络架构包含有 4 层,如果考虑到对于数据集的预处理,在视频帧序列和 3D 卷积神经网络之间还应该设置一个数据处理层,用于从原始图像中提取特别的信息。图 4 展示了模型的 4 层,序列中的每帧图像依序进入 3D 卷积神经网络层,然后长短期记忆神经网络接受卷积结果并且每个时间单位的隐藏层输出被传递给下一层[9]。一般的,需要在长短期记忆神经网络后面接一个全连接层或者聚合层,聚合层通常对所有的输入值进行平均操作,综合各个时间输出对于分类结果的影响[10-11]。最后 Softmax 分类器被用来对人体行为进行分类。

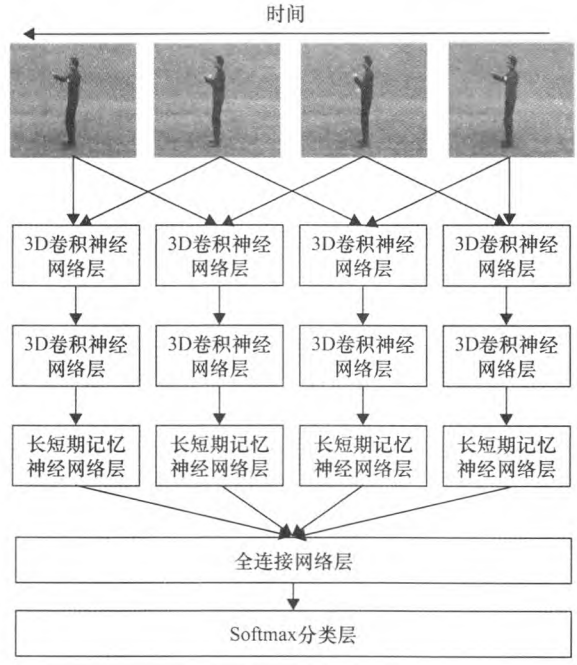

图 4　模型的网络结构

2　对照模型

2.1　长效递归卷积模型

LRCN 是由 Jeff Donahue 等人提出,用于视频识别和描述的深度学习模型[5,12]。其基本思想在于利用 CNN 抽象出单张图像的特征,输入至 LSTM 中学习时间关联特性。模型的每一个输入为可视化量 V_t(序列中的图像或者视频中的一帧),通过特征提取函数 $\varphi_V(v_t)$得到一个固定长度的特征向量 $\varphi_t \in \mathbf{R}^d$,利用序列模型学习特征向量集合$\{\varnothing_1,\varnothing_2,\cdots,\varnothing_T\}$。

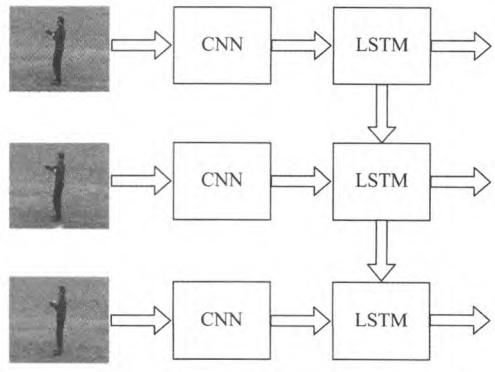

图 5　长效递归卷积神经网络的工作方式

2.2 时空卷积网络

时空域卷积神经网络(SCNN, spatiotemporal convolutional neural network)在文献[13]中被提出,模型认为三维行为信息被转换为相邻帧间二维图像变化关系,通过处理二维图像变化关系间接地学习三维行为信息[14]。时空域卷积神经网络有 5 层网络结构,对于输入图像矩阵 $\boldsymbol{X}$,$\boldsymbol{Y}$,存在 $\boldsymbol{Y}=\boldsymbol{L}\boldsymbol{X}$ 的变换关系。C 层为卷积层,存在 4 组卷积核 F_x,F_y,$\overline{F}_x$,$\overline{F}_y$,组成 2 对傅里叶函数对。M 层为乘积层,将 C 层得到的特征映射计算乘积;S 层为加法层,对 M 层输出特征映射积相加;P 层为池化层,减小微变化对于特征提取的影响。

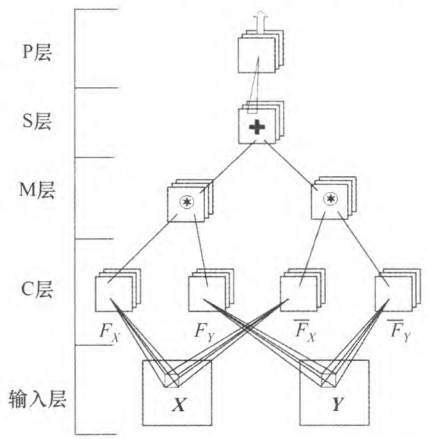

图 6 时空域卷积神经网络的网络结构[7]

3 实验测试

3.1 数据集

本文使用 KTH 数据集作为测试数据,它由固定摄像机采集的 600 个动作视频组成,视频中每帧都是一张 160 像素 ×120 像素的图像,图像背景均匀。

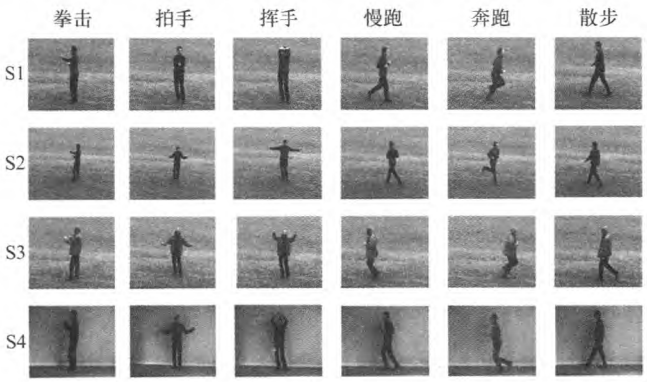

图 7 KTH 数据集分类

数据集共 600 个 AVI 视频,每个视频 25 f/s,总共包括 25 名实验对象,分属 4 类情景:户外 S1,户外(远近尺度变化)S2,户外(不同衣着)S3,室内 S4;6 类行为:拳击、拍手、挥手、慢跑、奔跑和散步。

本文选择 8 名实验对象作为训练数据集,8 名实验对象作为验证数据集,9 名实验对象作为测试数据集。其中,验证数据集用于在训练阶段进行测试,随着训练数据集不断迭代,每次迭代测试一次验证数据集,用于防止过拟合,防止导致模型对于训练数据集分类很好,对其他数据集效果很差。测试数据集在模型训练完成之后的测试阶段测试。视频保持 25 f/s 帧率存储为图像序列,每张图像可以视为 RGB 三通道矩阵。图像序列以 50 帧大小做分割,每帧图像大小压缩为 60 ×80,每个样本可以用一个 50 ×60 ×80 ×3 的高维数组表示。

对数据集进行预处理用于神经网络训练,由于原始数据为彩色图像,色彩通道不存在平稳特性,本文将像素值做简单缩放归一化到[0,1]区间,然后进行白化处理,利用 ZCA 不降低图像特征维度的特性,选用 ZCA 白化图像数据集。

3.2 实验结果

本文利用基于 Python 的深度学习库 Keras 在 GPU 并行加速环境下进行实验。3 种深度网络都后接一个 Dropout 层用于消除过拟合现象,输出通过 Softmax 层用于分类。实验环境如下:

① Intel i3 2.4 GHz 2 Cores;

② NVIDIA GeForce GTX 960(1024 个 CUDA 处理核心);

③ 8 GB 内存;

④ Ubuntu 14.04 ×64。

表 1 通过混淆矩阵对模型的识别结果进行了可视化,从表中可以看出,部分分类(拳击、慢跑和奔跑)模型识别错误的情况较多,而挥手、拍手和散步的识别率很高,对于 KTH 数据集的平均识别率达到了 93.7%。从原始的视频中可以发现慢跑和奔跑的区分度不大,数据本身的识别难度很高。因此,模型仍然具有很好的鲁棒性能。

表 1 模型识别结果

		模型预测					
		散步	慢跑	奔跑	拳击	挥手	拍手
实际分类	散步	99	1	—	—	—	—
	慢跑	4	88	6	1	—	1
	奔跑	5	4	90	—	—	—
	拳击	—	5	—	91	3	1
	挥手	—	—	—	1	95	4
	拍手	—	—	—	—	4	96

图 8 显示了 3 种深度神经网络对于 KTH 数据集识别率的变化情况,列所代表的是识别率,行代表 epoch 迭代次数。在整个训练过程中,3 条曲线代表的识别率没有出现下降情况,这表明训练未发生严重的过拟合现象。从结果中可以看出时空域卷积网络在训练

量较小时学习速度较快,而当迭代次数逐渐增大时,可以由图看出本文提出的融合模型识别率要高于另外两种网络结构。

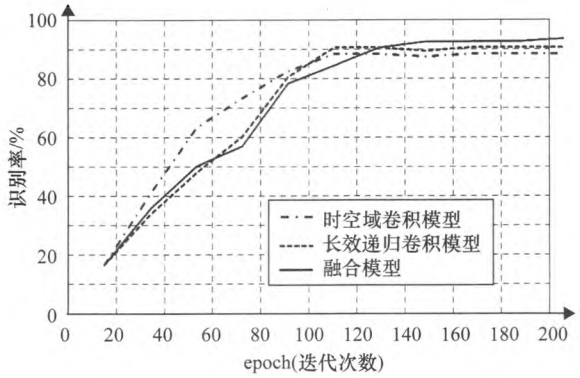

图 8　3 种模型对 KTH 数据集的识别情况

4　结束语

本文提出了 3D 卷积神经网络和长短期记忆神经网络的融合模型。在 KTH 数据集上的识别率达到 93.7%。相比较长效递归卷积模型和时空卷积网络,模型能够提取每帧图像的特征和帧与帧之间的关联特征,在采样频率低或持续时间长的动作序列上,识别率提升明显。整个模型基于深度网络,无需先验经验,具有良好的泛化性能。

本文没有将更显著的特征加入模型中,如果运用一些已经被验证高效的特征(如 3D-SHIFT、Cuboids 描述子),模型可以进一步改进。未来,会研究这些高效特征对 3D 卷积神经网络和长短期记忆神经网络组合识别率的影响。

参考文献:

[1] Davis J W, Bobick A F. The representation and recognition of action using temporal templates[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2000:928 - 934.

[2] Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1): 221 - 231.

[3] Wernicke S. A Faster Algorithm for Detecting Network Motifs [M]//Algorithms in Bioinformatics. Berlin Heidelberg: Springer, 2005:165 - 177.

[4] Graves A. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735 - 80.

[5] Donahue J, Hendricks L A, Guadarrama S, et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description[M]// AB Initto Calculation of the Structures and Properties of Molecules. Elsevier, 1988:85 - 91.

[6] 郝宗波,桑楠,吴杰,等. 一种基于 3D 卷积神经网络的行为识别方法:中国,104281853A[P]. 2015.

[7] 吴杰. 基于卷积神经网络的行为识别研究[D]. 成都:电子科技大学,2015.

[8] 张亮,黄曙光,石昭祥,等. 基于 LSTM 型 RNN 的 CAPTCHA 识别方法[J]. 模式识别与人工智能,2011,24(1):40 - 47.

[9] Molchanov P, Gupta S, Kim K, et al. Hand gesture recognition with 3D convolutional neural networks[C]//IEEE Computer Vision and Pattern Recognition Workshops. 2015.

[10] Sundermeyer M, Ney H, Schluter R. From feedforward to recurrent LSTM neural networks for language modeling[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 23(3):517 - 529.

[11] Cai M, Liu J. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition[J]. Speech Communication, 2015, 77(C):53 - 64.

[12] Baccouche M, Mamalet F, Wolf C, et al. Action classification in soccer videos with long short-term memory recurrent neural networks [C]//20th International Conference on Aritificial Neural Networks. 2010:154 - 159.

[13] 刘琮,许维胜,吴启迪. 时空域深度卷积神经网络及其在行为识别上的应用[J]. 计算机科学,2015(7):245 - 249.

[14] Roland M. Learning to relate images[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8): 1829 - 1846.

□