计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190500154

基于卷积神经网络的半监督手术视频流程识别

祁宝莲[1,3]　钟坤华[1,2,3]　陈芋文[1,2,3]
1 中科院成都计算所　成都 610041
2 中国科学院重庆绿色智能技术研究院　重庆 400714
3 中国科学院大学　北京 100049
(qibaolian17@mails.ucas.ac.cn)

摘　要　实时鲁棒的开放性外科手术视频流程自动识别检测将是未来人工智能医疗手术室的核心组成部分,这一关键技术结合其他 AI(Artificial Intelligence)技术就可以帮助医护人员自动化、智能化地完成多项术中的常规活动。利用人工智能和计算机视觉的方法进行手术流程识别检测需要对大量的数据进行学习,为了训练这种方法,需要大量地标记手术视频数据,然而在医学领域,对外科手术视频数据的标记需要专家知识,收集足够数量的标记外科手术视频数据是困难且耗时的。因此,文中以腹腔镜胆囊切除术视频数据为研究对象,通过半监督学习方法卷积自编码器对视频进行空间特征提取,结合从同视频上下文中的一对视频帧进行时序特征提取,将非结构化的手术视频数据结构化,从而构建低层手术视频特征到高层外科手术流程语义之间的桥梁,以低代价实现对手术视频流程的智能化识别检测,高效判定手术流程进展。在开源数据集上的实验的结果表明,使用该模型 jacc 系数达到 71.3%,准确率为 86.6%,取得了较好的实验效果。

关键词:半监督;手术流程;卷积网络
中图法分类号　TP181

Semi-supervised Surgical Video Workflow Recognition Based on Convolution Neural Network

QI Bao-lian[1,3],ZHONG Kun-hua[1,2,3] and CHEN Yu-wen[1,2,3]
1 Chengdu Computing Institute of the Chinese Academy of Sciences,Chengdu 610041,China
2 Chongqing Institute of Green and Intelligent Technology,Chongqing 400714,China
3 University of Chinese Academy of Sciences,Beijing 100049,China

Abstract　The real-time and robust open surgery workflow automatic detection will be the core component of the future artificial intelligent medical operation room. The key technology combined with other artificial intelligence technologies can help medical staff to automatically and intelligently complete a number of routine activities in the operation. However,the use of artificial intelligence and computer vision for surgical workflow recognition requires a large amount of data to be learned. In order to train this method,a large amount of labeled surgical video data is required. However,in the medical field,the labeling of surgical video data requires expert knowledge,and collecting enough numbers of marked surgical video data is difficult and time-consuming. Therefore,in this paper,the video data of laparoscopic cholecystectomy data is taken as the research object,the video spatial feature extraction is carried out by convolution self-encoder with semi-supervised learning method,and combined with a pair of video frames in the context of the same video for sequential feature extraction. The unstructured surgical video data is structured to build a bridge between the video characteristics of low-level surgery and the semantics of high-level surgical procedures,trying to realize the intelligent recognition of the surgical workflow at a low cost,and effectively determining the progress of the surgical workflow. Finally,the jaccard coefficient of the proposed algorithm in this paper on a public dataset is 71.3% and the accuracy is 86.6%,achieving good experimental results.

Keywords　Semi-supervised,Surgical workflow,CNN

1　引言

随着现代精准外科范式的建立,手术种类与形式越来越多,手术过程也越来越复杂,利用术前和术中的多种影像全面助力外科手术是大势所趋[1]。鉴于此,围绕外科手术及其相关活动的智能化已经成为当前 AI 医疗的研究热点[2]。然而,实现真正意义上的手术及其相关活动智能化的核心前提和挑战在于计算机对外科手术视频流程的识别和理解。通过自动手术流程检测,来识别手术视频帧当前所处的阶段,能解决当前非常耗时的手术视频数据库的自动化索引问题。例

基金项目:国家重点研发计划项目(2018YFC0116704);重庆市技术创新与应用发展专项面上项目(cstc2019jscx-msxmX0237)
This work was supported by the National Key Research & Development Plan of China (2018YFC0116704) and Chongqing Technology Innovation and Application Development Project(cstc2019jscx-msxmX0237).
通信作者:陈芋文(chenyuwen@cigit.ac.cn)

如，外科工作人员搜索特定的视频片段用于讨论或学习，需要手动定位视频所处位置。而自动化识别手术流程可以减轻他们执行简单耗时的任务。其中，实时手术流程识别能够提高手术过程中病人的安全性，根据手术内容识别可以提前预警手术过程中可能出现的并发症[3]，从而更好地辅助外科医生的术中工作。同时，手术流程分析能够预测手术剩余时长[4]，为手术室外的临床工作人员提供手术进展情况，方便其后续工作的安排及优化医院管理工作[5]。因此，对手术视频进行有效分析在大数据智能时代极其重要。

手术流程是一个具有很强逻辑性的动态过程，要求计算机根据人类的思维和感知来分析手术流程。计算机对其进行智能分解、识别是一个情景感知的过程，其中如何跨越低级像素特征和高级语义特征之间的鸿沟是手术视频流程理解中最重要的一步，也是计算机视觉视频理解领域颇具挑战的一个研究课题[6]。文献[7]首次提出了手术流程分析，将腹腔镜胆囊切除术分成 6 个阶段，通过强化学习的方法对腹腔镜胆囊切除术视频进行特征提取，然后通过支持向量机（Support Vector Machine，SVM）对提取的特征进行手术阶段的识别理解，识别的准确度达到 50%。Blum 等[8]同样基于内窥镜腹腔胆囊切除术进行 14 个阶段的识别，对手术视频使用梯度大小、直方图和颜色值等简单的特征，然后用典型相关法分析（Canonical Correlation Analysis，CCA）和主成分分析（Principal Component Analysis，PCA）进行降维，然后基于隐马尔可夫模型（Hidden Markov Model，HMM）、动态时间规整算法（Dynamic Time Warping，DTW）等方法进行阶段检测。DTW 算法产生了最好的性能检测准确率 76.8%。Dergachyova 等[9]融合手术器械数据对手术流程进行检测识别。该方法首先采用对手术进行过程建模，然后对视觉和手术器械进行特征提取，接着用 AdaBoost 概括特征对手术阶段进行分类，最后用隐马尔可夫模型生成最终决策。基于视觉特征，该算法的精度接近 68%，融合手术器械特征的准确度接近 90%。Twinanda 等[10]提出了 Endonet 架构，一个基于卷积神经网络的架构，对手术流程进行在线和离线学习和识别。该方法仍基于腹腔镜胆囊切除术在两大数据集（Cholec80 和 EndoVis）上进行验证，达到了较好的性能。后来，Jin 等[3]提出了一种由卷积神经网络（CNN）和长短时记忆网络（LSTM）组成的 SV-RCNet 来识别手术视频流程。该网络利用 CNN 从视频帧中提取视觉特征，利用 LSTM 网络提取手术视频中的时序信息。Loukas 等[11]研究了 4 种架构（Alexnet，VGG19，GoogleNet 和 ResNet101），通过迁移学习的方法来提取特征，同时运用视觉显著性方法选择图像中信息最显著区域进行学习。

利用人工智能和计算机视觉的方法进行手术流程识别检测需要对大量的数据进行学习，为了训练这种方法，需要大量的手术视频标记数据，然而在医学领域，对外科手术视频数据的标记需要专家知识，收集足够数量的标记外科手术视频数据是困难且耗时的。因此，本文提出了一种更符合现实情境的方法，即用少量标记的方法来进行手术流程的识别。本文基于深度学习理论中的卷积神经，采用半监督的方式对腹腔镜胆囊切除术进行流程自动化的识别和研究。

2 半监督手术流程识别方法

深度学习解决了特征自动选择和高层次特征提取的难题，是近年来机器学习领域研究的新方向[12]。本文的无监督视频特征学习（Unsupervised Feature Learning，UFL）基于深度学习理论（Deep Learning，DL）中的卷积神经网络模型，通过卷积自编码网络[13]和卷积时序判别网络预训练从无标号数据中自动提取特征，并将这些特征用于识别任务。

2.1 无监督卷积自编码器的特征提取

卷积自编码器经过无监督逐层贪心预训练和系统性参数优化的多层非线性卷积网络，从无类标数据中提取高维复杂输入数据的分层特征。计算流程如下（见图 1）：1）将手术视频数据输入卷积自编码器的编码网络（Encoder），所有局部特征共享权值矩阵，隐含层完整保存受局部空间限制的边缘特征，利用重要的局部特征重构原始数据；2）将经过编码器的特征输入解码网络（Decoder），对信息进行解码以复现原图。解码器用影像块的线性组合来拟合原有图像，实现特征的分层提取，显著提升了图像识别的速度和准确率；3）最后用损失函数来进行计算，重构图像并确定网络参数。

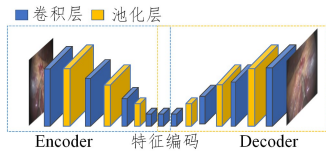

图 1 卷积自编码网络

Fig. 1 Convolutional self-encoder network

手术流程视频图片为 x，网络中有 k 个卷积核，则 k 阶特征映射可表示为：

$$h_k=\sigma(x*\boldsymbol{w}_k+b_k)$$

其中，σ 为激活函数，$\boldsymbol{w}_k$ 为权值矩阵，b_k 为偏置，$*$ 表示卷积，重构函数为：

$$\bar{x}=\sigma\Big(\sum_{k\subset H}h_k*\boldsymbol{w}_k+b_k\Big)$$

其中，H 是隐藏特征映射集，$\boldsymbol{w}_k$ 为权值矩阵，b_k 为偏置。

将输入的样本和最终利用特征重构得出的结果进行欧几里得距离比较，通过反向传播算法（BP）进行优化，得到完整的卷积自编码器。

$$E=\frac{1}{2n}\sum_{i=1}^{n}(x_i-\bar{x}_i)^2$$

其中，n 为样本数据量，通过 SGD 算法训练网络，求解网络权值。

2.2 无监督卷积时序判别网络的特征提取

外科手术是一个具有很强逻辑性的动态过程，具有相对固定的逻辑顺序。如图 2 所示，在同一手术的两帧中，两帧最可能的相对顺序是哪一帧？即皮肤切割阶段必然先于皮瓣分离阶段（左边先于右边）。考虑到这种时序特征，通过利用手术流程的自然特性，设计卷积时序判别网络对手术视频进行时序排序任务学习，从而减少模型训练所需要的样本数量，具体方法如图 2 所示。

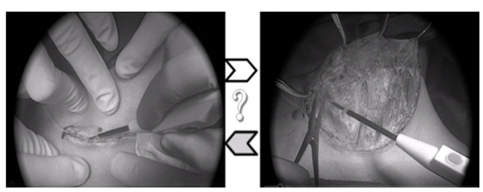

图 2 手术帧时序上的相对顺序示意图（左边先于右边）

Fig. 2 Schematic diagram of relatively temporal order of operation frames (the left before the right)

用预先训练的 Encoder 网络对胆囊切除术视频帧按时间顺序进行排序学习，以每秒一帧的间隔来提取来自同一视频的一对帧被馈送到已训练的 Encoder 网络的两个输入层，每帧都由 Encoder 网络处理，两条链中的相应层分享权重，然后将两个 Encoder 网络的输出连接起来，最后使用两个全连接层进行处理，如果第一帧先于第二帧则输出 1，否则输出 0，进行无监督的二次特征提取，如图 3 所示。

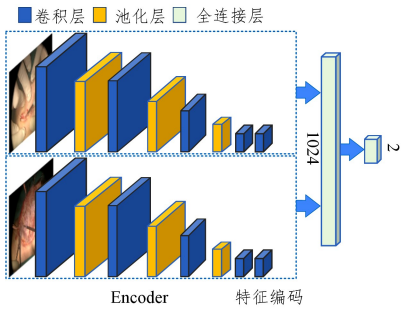

图 3 卷积时序判断网络

Fig. 3 Convolutional temporal determine network

两个视频帧分别提取的特征为 z_1 和 z_2，并通过全连接层连接两个特征，最后计算 Logistic 损失函数：

$$z=z_1+z_2$$

$$h_\theta(z)=\frac{1}{1+e^{-\theta z}}$$

损失函数的定义如下：

$$cost(h_\theta(z_i),y_i)=y_i\log(h_\theta(z_i))+(1-y_i)\log(1-h_\theta(z_i))$$

$$J(\theta)=\frac{1}{n}\sum_{i=1}^{n}cost(h_\theta(z_i),y_i)$$

其中，n 为样本量，训练时固定 Encoder 网络的权值，训练时序网络最后一层权值。

2.3 半监督自动化流程识别

基于卷积自编码、时序判别网络无监督的方式训练网络模型，将卷积时序判别网络的损失函数调整为 softmax 函数，对手术视频数据进行半监督的学习，微调整体网络，对手术进行流程识别，如图 4 所示。

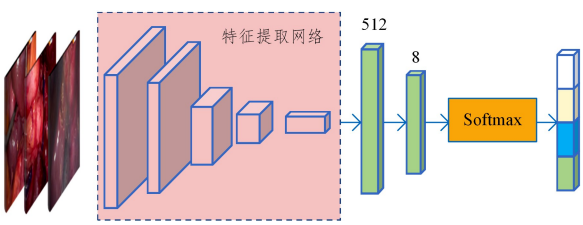

图 4 手术流程识别示意图

Fig. 4 Schematic diagram of surgical workflow recognition

其中，softmax 损失函数的定义如下：

假设手术视频帧数据集 $D=\{(x_1,y_1),\cdots,(x_m,y_m)\}$，其中 x_m 表示手术帧特征，k 表示向量维数，y_m 表示该帧对应的手术阶段号，对于给定手术视频的第 i 帧，定义每个流程阶段的概率为：

$$p(y_i=j\mid x_i;\theta)=\frac{e^{\theta_j^{T}x_{i,j}}}{\sum_{l=1}^{k}e^{\theta_j^{T}x_{i,j}}}$$

然后通过最小化预测概率和 Groundtruth 真实类别的 KL 散度，得到如下损失函数：

$$J=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}\bar{p}_{ij}\log(p_{ij})$$

其中，p_{ij} 表示第 i 帧属于第 j 个流程阶段的概率，$\bar{p}_{ij}$ 为真实类别概率。

3 实验验证及结果

3.1 训练数据集

我们采用手术流程检测竞赛的 m2cai16-workflow 数据集进行实验验证。该数据集包含 41 个腹腔胆囊切除术视频，从特拉斯堡/ IRCAD 大学医院(斯特拉斯堡，法国)和慕尼黑工业大学附属(慕尼黑，德国)以帧率 25 fps 采集获得。其分为两部分：训练子集(包含 27 个视频)和测试子集(14 个视频)。将手术分为 8 个流程阶段进行注释标记，详细信息如表 1 所列。

表 1 手术流程阶段

Table 1 Surgical procedure stage

阶段号	阶段名称
1	trocar placement
2	preparation
3	calot triangle dissection
4	clipping and cutting
5	gallbladder dissection
6	galbladder packaging
7	cleaning and coagulation
8	gallbladder retraction

3.2 评估指标

本文用下面的指标来评估手术流程识别方法的性能。

1)召回率：某流程阶段正确样本占全部正样本的百分比。

2)准确率：所有流程阶段预测正确的百分比。

3)Jaccard 系数：某流程阶段预测值与真实值交集与并集的比值。

对于所有手术视频，我们计算所有阶段的平均准确率和平均 jaccard 系数。

3.3 模型训练

基于计算平台编程实验验证卷积自编码网络和卷积时序判别网络的拓扑结构。在医学图像处理领域，U-NET 网络[14]是基于全卷积网络的一个语义分割网络，适合用来做医学图像的特征提取。以 U-NET 网络为基础结构，将 U-NET 网络的 Encoder 网络改进为卷积编码的 Encoder 网络。此外，将通道数降为原始网络的一半，去掉原始网络的快捷连接，具体结构如图 4 所示。

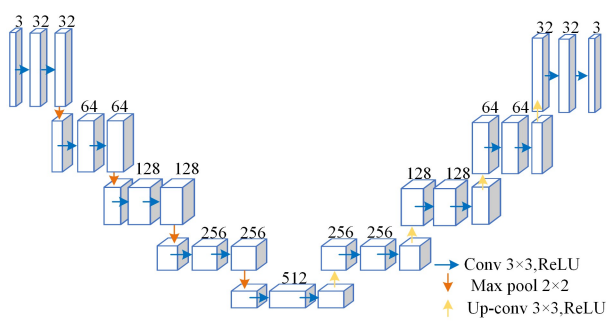

图 5 卷积自编码器设计图

Fig. 5 Design diagram of convolution self-encoder

通过无监督训练后，Encoder 网络连接两层全连接层，如图 4 所示，采样手术视频分辨率为 256×256，将帧中 RGB 通道像素值映射到[−0.5,0.5]，整个网络有 18 个卷积层，4 次下采样，4 次上采样，通过 SGD 优化算法来训练网络以求解参数。初始学习率 λ 被设置为 $10-10^{-3}$，动量设置为 0.9，批

量大小为 256，为防止过拟合采用了 L1 和 L2 正则化。

3.4 实验

我们通过 3 组实验来验证无监督特征提取的有效性。

1）训练卷积自编码器网络（CAE）后进行手术流程分割。

2）训练卷积时序判别网络（CTD）后进行手术流程分割。

3）融合卷积自编码＋卷积时序判别网络后进行手术流程分割。

3 组实验分别进行，计算各个手术流程阶段的评估指标，最后以平均准确率和平均 jaccard 系数来评估模型的效果，结果如表 2－表 4 所列。

表 2 CAE 网络

Table 2 CAE network

阶段	Jacc	Prec	Rec
trocar placement	53.4	70.0	73.0
preparation	39.9	56.8	80.9
CalotTriangleDissection	62.0	90.5	69.9
clipping and cutting	53.3	73.4	72.9
gallbladder dissection	53.1	67.3	80.6
galbladder packaging	61.1	95.8	70.4
CleaningCoagulation	51.1	66.4	69.8
gallbladder retraction	47.8	66.5	64.4
平均值	**52.7**	**68.3**	—

表 3 CTD 网络

Table 3 CTD network

阶段	Jacc	Prec	Rec
trocar placement	68.2	80.4	83.2
preparation	54.3	67.4	85.3
CalotTriangleDissection	70.5	95.2	74.3
clipping and cutting	64.2	81.4	80.3
gallbladder dissection	71.4	75.3	74.2
galbladder packaging	66.7	97.3	74.5
CleaningCoagulation	53.9	69.4	67.3
gallbladder retraction	72.4	78.9	76.3
平均值	**65.2**	**80.7**	—

表 4 CAE＋CTD 网络

Table 4 Network of CAE&CTD

阶段	Jacc	Prec	Rec
trocar placement	70.4	85.2	80.3
preparation	72.2	89.1	81.2
CalotTriangleDissection	75.3	97.8	72.1
clipping and cutting	70.3	85.2	75.2
gallbladder dissection	69.5	87.2	76.4
galbladder packaging	80.3	98.0	70.5
CleaningCoagulation	60.1	70.2	69.8
gallbladder retraction	72.5	80.2	77.5
平均值	**71.3**	**86.6**	—

结束语 本文提出了一种无监督的手术视频流程识别的方法，解决了在自动手术流程识别中常见的标签数据不足的问题。本文首先用卷积自编码器对视频进行空间特征提取，然后从同视频上下文中提取一对视频帧进行时序排序任务学习，并提取时序特征，最后通过多分类识别手术流程阶段。本文在公开的腹腔胆囊切除术视频数据集上验证了提出的网络，实验取得了不错的效果，其中，jacc 系数为 71.3，平均准确率为 86.6%。本文研究展现了无监督学习方法在医疗手术流程识别中具有的强大的应用潜力，同时也表明自动手术流程识别方法在临床实践中具有极其重要的意义。在将来的工作中，作者将尝试使用生成对抗网络来进行手术流程无监督的学习，从而生成手术流程图片，以解决医疗手术视频数据标记问题。

参 考 文 献

[1] 徐大华. 高科技引领微创外科发展[J]. 科技导报，2017，35(11)：69-70.

[2] 曹晖. 人工智能医疗给外科医生带来的挑战、机遇与思考[J]. 中国实用外科杂志，2017(12)：387-388.

[3] JIN Y，DOU Q，CHEN H，et al. SV-RCNet：Workflow recognition from surgical videos using recurrent convolutional network [J]. IEEE Trans. Med. Imaging，2018，37(5)：1114-1126.

[4] TWINANDA A，YENGERA G，MUTTER D. RSDNet：Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations[J]. arXiv：1802. 03243v2，2018.

[5] LOUKAS C. Video content analysis of surgical procedures[J]. Surgical Endoscopy，2018，32(2)：553-568.

[6] LI X，URICCHIO T，BALLAN L，et al. Socializing the Semantic Gap：A Comparative Survey on Image Tag Assignment，Refinement，and Retrieval[J]. Acm Computing Surveys，2016，49(1)：14.

[7] KLANK U，PADOY N，FEUSSNER H. Navab N (2008) Automatic feature generation in endoscopic images[J]. Int J Comput Assist Radiol Surg 3：331-339.

[8] BLUM T，FEUSSNER H，NAVAB N. Modeling and segmentation of surgical workflow from laparoscopic video[J]. Lect Notes Comput Sci，2010，6363：400-407.

[9] DERGACHYOVA O，BOUGET D，HUAULMÉ A，et al. Automatic data-driven real-time segmentation and recognition of surgical workflow[J]. Int J Comput Assist Radiol Surg，2016，11：1081-1089.

[10] TWINANDA A P，SHEHATA S，MUTTER D，et al. EndoNet：a deep architecture for recognition tasks on laparoscopic videos [J]. IEEE Trans Med Imaging，2017，36：86-97.

[11] LOUKAS C. Surgical Phase Recognition of Short Video Shots Based on Temporal Modeling of Deep Features[J]. arXiv：1807. 07853，2018.

[12] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436.

[13] GENG C，SONG J X. Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder [C]// International Conference on Computer Sciences and Automation Engineering. Atlantis Press，2016.

[14] RONNEBERGER O，FISCHER P，BROX T. U-Net：Convolutional Networks for Biomedical Image Segmentation[M]// Medical Image Computing and Computer-Assisted Intervention－MICCAI 2015. Springer International Publishing，2015：234-241.

QI Bao-lian，postgraduate. Her main research interests include video analysis，surgical workflow recognition，and artificial intelligence for healthcare.

CHEN Yu-wen，doctorial student. His main research interests include automated reasoning and programming，computer vision and artificial intelligent for healthcare.