基于深度学习的头部姿态估计

贺飞翔,赵启军
(四川大学 视觉合成图形图像技术国防重点学科实验室,四川 成都 610065)

摘 要:头部姿态估计在人工智能、模式识别及人机智能交互等领域应用广泛。好的头部姿态估计算法应对光照、噪声、身份、遮挡等因素时鲁棒性较好,但目前为止如何提高姿态估计的精确度与鲁棒性依然是计算机视觉领域的一大挑战。提出了一种基于深度学习进行头部姿态估计的方法。利用深度学习强大的学习能力,对输入的人脸图像进行一系列的非线性操作,逐层提取图像中抽象的特征,然后利用提取的特征进行分类。此类特征在姿态上具有较大的差异性,同时对光照、身份、遮挡等因素鲁棒。在 CAS-PEAL 数据集上对该方法进行了评估实验。实验结果表明,该方法有效地提高了姿态估计的准确性。

关键词:头部姿态估计;深度学习;提取特征;分类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2016)11-0001-04
doi:10.3969/j.issn.1673-629X.2016.11.001

Head Pose Estimation Based on Deep Learning

HE Fei-xiang, ZHAO Qi-jun
(National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China)

Abstract: Head pose estimation has been widely used in the field of artificial intelligence, pattern recognition and intelligent human-computer interaction and so on. Good head pose estimation algorithm should deal with light, noise, identity, shelter and other factors robustly, but so far how to improve the accuracy and robustness of attitude estimation remains a major challenge in the field of computer vision. A method based on deep learning for pose estimation is presented. Deep learning with a strong learning ability, it can extract high-level image features of the input image by through a series of non-linear operation, then classifying the input image using the extracted feature. Such characteristics have greater differences in pose, while they are robust of light, identity, occlusion and other factors. The proposed head pose estimation is evaluated on the CAS-PEAL data set. Experimental results show that this method is effective to improve the accuracy of pose estimation.

Key words: head pose estimation; deep learning; extracting feature; classification

1 概 述

头部姿态是研究人类行为和注意力的关键,在人际交往中,扮演着非常重要的角色。头部姿态的改变也包含丰富的信息,例如同意、反对、理解、迷惑、惊喜等。此外,头部姿态还是包括人脸识别、表情识别、视线估计在内的许多智能系统在非约束条件下进行身份识别与行为预测所需要的关键信息。因此,头部姿态估计是计算机视觉与模式识别领域一个非常重要的应用,其算法研究的意义非常大。

在计算机视觉领域,头部姿态估计[1]是指计算机通过对输入图像或者视频序列的分析、预测,确定人的头部在三维空间(相对于摄像机)中的位置及姿态参量。通常说来,假设头部姿态估计是一个刚体变换,存在 pitch,yaw,roll 三个方向自由度,如图 1 所示。由于受非约束坏境中的投影几何形变、背影光照变化、前景遮挡问题和低分辨率等因素的影响,使得头部姿态的多自由度估计一直是一个富有挑战性的领域。

针对人脸头部姿态估计的算法,主要分为基于模型的方法和基于人脸表观的方法。其中,基于模型的方法[2-4]主要是利用若干脸部特征点构成的模型,通过提取不同姿态下模型的差异预测头部姿态。此类方法实现简单、计算高效准确、易于理解,但强烈依赖特

收稿日期:2016-01-29 **修回日期**:2016-05-18 **网络出版时间**:2016-10-24
基金项目:国家自然科学基金资助项目(61202160,61202161);科技部重大仪器专项(2013YQ49087904)
作者简介:贺飞翔(1992-),男,硕士研究生,研究方向为生物特征识别;赵启军,副教授,硕士研究生导师,研究方向为生物特征识别。

征点定位的准确程度,而准确的特征点检测在姿态变化较大时仍然是一个急待解决的挑战。基于表观学习[5-10]的方法是通过大量的训练数据直接学习图像与头部姿态之间的映射关系。与基于模型匹配方法相比,其主要优点是提取基于表观的特征不依赖特征点的位置,具有较高的鲁棒性与估计精确度。文中研究的方法属于基于表观学习的方法。

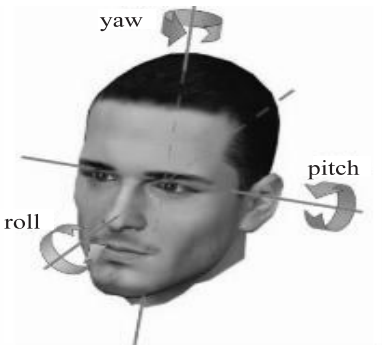

图 1　人脸头部姿态三个旋转方向

基于图像表观的学习方法解决姿态的问题通常可以看作是分类问题或者是回归问题。此类算法的核心主要分为两部分:第一部分是提取图片中与头部姿态变化紧密相关的特征,此类特征对人的身份、表情、光照等无关因素鲁棒;第二部分是通过某种分类算法,对提取的特征进行分类或回归对姿态角度进行估计。

深度学习是机器学习中一个新的领域。从 2006 年开始,深度学习在语音识别、计算机视觉(包括人脸识别、特征点检测、人脸检测等)、自然语言处理以及信息检索等领域性能优异。深度学习可通过学习一种深层非线性网络结构,实现复杂函数逼近,其特有的层次结构能够对数据局部特征进行多层次抽象化的学习与表达。文中主要是利用深度学习强大的学习能力,学习输入图片中与对象的身份、光照、表情等因素无关,且仅与姿态有关的特征,然后通过分类,用以解决头部姿态估计的问题。

2　基于卷积网络头部姿态估计

2.1　基于 CNN 的深度学习网络

文中采用的深度网络结构模型主要包括 2 个卷积层(含 2 个采样层),后接 1 个全连接层和 soft-max 输出层。如图 2 所示(图中外面的大立方体的长、宽、高分别表示每一层特征图的个数与特征图的维度,里面的小立方体和正方形分别表示卷积过程中卷积核的尺寸与下采样过程中采样矩形框的尺寸,最后两层是全连接中神经元的个数),输入图片 x_0 是尺度大小归一化至 32 × 32 的灰度图像。图像输入到网络结构,逐层对输入图片进行卷积与池化采样,提取抽象的特征,通过 soft-max 对提取的抽象特征分类,网络的最终输出为输入图片的头部姿态。当输入图片的尺寸发生变化时,网络结构中每一层特征谱的高与宽都会发生相应的变化。在提取图像特征的过程中,特征逐渐抽象化,特征的维度逐渐下降,形成更加简洁抽象且具有高度区分性的特征,从而能够正确分类出输入图片中头部姿态所属类别。

在卷积阶段,利用卷积核对特征图进行卷积操作,加强原信号信息,并且降低图片噪音。在卷积神经网络中,每个卷积核能够提取输入特征图中所有位置上的某一特定特征,每一个卷积滤波器共享相同的参数,包括相同的权值矩阵与偏置项,从而实现同一个输入特征图上的权值共享[11]。权值共享的优点是在对图片提取特征时不用考虑局部特征权重的差异(比如鼻子、眼睛、嘴巴),使要学习的卷积神经网络模型的参数数量大大降低。

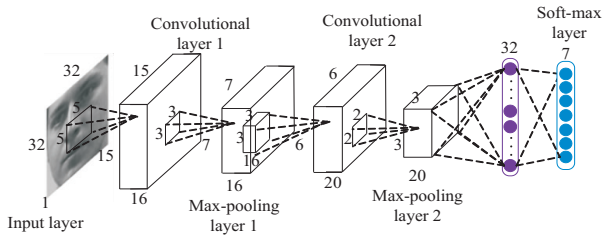

图 2　深度网络结构详图

为了提取能够预测输入图片中人脸头部偏转角度的多个特征,需要使用不同的卷积核进行卷积操作。卷积运算可表示如下:

$$y^j = \max(0, b^j + \sum_i w^{ij} * x^i) \tag{1}$$

其中,x^i 与 y^j 分别表示第 i 个输入特征图与第 j 个输出特征图;w^{ij} 是第 i 个输入特征图与第 j 个输出特征图之间的卷积核;$*$ 表示卷积;b^j 是第 j 个输出图的偏置项。

为了避免线性模型表达能力不够,通常需要对卷积过后得到的特征图进行非线性化操作,防止过拟合。常用的非线性函数主要有 sigmoid、tanh、ReLU 等。文中对隐层神经元使用不饱和非线性函数 ReLU。

下采样主要是实现特征的降维。由于图像局部相关性原理,通过对图像进行下采样,在保留图像有用信息的同时降低了特征图的维度。下采样阶段主要是对单个特征图进行操作,主要有平均池化下采样与最大池化下采样。平均池化下采样是取邻域中的平均值作为输出,最大池化下采样是取邻域中的最大值作为输出。文中采用的是最大池化下采样,过程可表示为:

$$y^i_{j,k} = \max_{0 \leqslant m,n \leqslant s} \{x_{j\cdot s} + m, k \cdot s + n^i\} \tag{2}$$

其中,y^j 表示下采样过程中的第 i 个输出谱,其中的每一个神经元是从第 i 个输入谱中 $s \times s$ 局部区域采样得到的;m 与 n 分别表示下采样框移动的步长。

SoftMax 回归是在逻辑回归的基础上扩张而来的,主要是为了解决多分类问题,是有监督的学习算法。网络的最后一层是 SoftMax 函数,与深度学习结合使用,用来区分输入图片的角度类别。

$$y_i = \frac{\exp(y_i')}{\sum_{j=1}^{n} \exp(y_j')} \tag{3}$$

其中,$y_j' = \sum_{i=1}^{128} x_i \cdot w_{i,j} + b_j$ 表示 128 个特征 x_i 的线性组合作为第 j 个神经元的输入;y_i 表示第 j 个神经元的输出。

因此整个网络的优化目标是最小化 $-\log y_t$。

2.2　网络的训练与测试

训练深度网络模型的本质就是获得构建网络的所有参数(包括权重与偏置),其训练的复杂程度与参数的数量正相关。

训练网络流程图如图 3 所示。

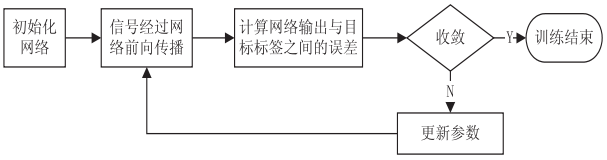

图 3　训练网络流程图

Step1:图片预处理。用双线性内插法将测试样本与训练样本归一化至尺寸为 32 × 32 的灰度图像。

Step2:将训练样本输入网络并前向传播,计算网络的输出与给定的目标标签之间的误差,判断是否有收敛。

Step3:若收敛,则训练结束;若不收敛,则误差反向传递,逐层更新参数,然后转到 Step2。

网络训练过程中,采用的是反向传递的方法以及有监督的训练方式。通过计算在每一层中误差对参数 W 的导数,从而更新参数,反复迭代使网络收敛。深度卷积神经网络可以看作是多个卷积模块串联在一起。每一个模块可以用公式 $y = f(X, W)$ 表示,则网络结构的抽象表示如图 4 所示。

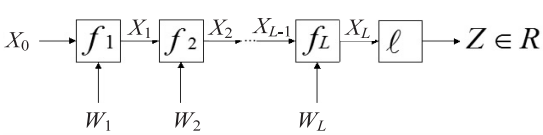

图 4　网络结构的抽象表示

整个网络的误差对每一层参数求导的公式为:

$$\frac{\mathrm{d}z}{\mathrm{d}(W_l)^{\mathrm{T}}} = \frac{\mathrm{d}z}{\mathrm{d}(X_L)^{\mathrm{T}}} \frac{\mathrm{d}X_L}{\mathrm{d}(X_L - 1)^{\mathrm{T}}} \cdots \frac{\mathrm{d}X_{l+1}}{\mathrm{d}(X_l)^{\mathrm{T}}} \frac{\mathrm{d}X_l}{\mathrm{d}(W_{l-1})^{\mathrm{T}}} \tag{4}$$

在网络测试阶段,将测试图片通过训练好的网络,网络的输出即为测试图片的角度类别标签。

3　实验与结果分析

3.1　实验数据库

文中实验主要是在 CAS-PEAL 数据集上进行。CAS-PEAL 是进行头部姿态估计常用的一个数据集。在该数据集上,头部姿态被划分为 7 个 yaw 方向上的离散角度{-45°,-30°,-15°,0°,15°,30°,45°}和 3 个 pitch 方向的离散角度{-30°,0°,30°}。对于每一张图片,使用一个人脸检测器[12]定位图片的人脸区域,截取人脸图片并将其归一化至 32 * 32。CAS-PEAL 中一些样本图片如图 5 所示。

图 5　CAS-PEAL 数据集中的样本图片

3.2　实验结果

在 CAS-PEAL 数据集中对象编号为 401 ~ 600 的子集上进行实验。该数据子集上共有 4 200(21 * 200)张图片,使用人脸检测器检测到的人脸图片共 4 166 张。使用三折交叉验证,将实验数据集分成三个数据子集,其中一个数据子集用来测试,剩下的两个用来训练。通过这样的方式,保证所有训练图片与测试图片不交叉。重复三次实验使每一个子集都参与测试,实验结果为三次测试结果的平均值。在该实验数据集上,实验结果如表 1 所示。其中,VoD 与 kVoD 使用的是另一个人脸检测器[13]。

表 1　CAS-PEAL 数据集上几种方法的平均绝对误差(MAE)比较

Method	Yaw error (MAE)
VoD[14]	4.6
kVoD[14]	4.1
DNNP	1.36

在 CAS-PEAL 数据集上,分别取编号为 401 ~ 600、201 ~ 600、201 ~ 800、201 ~ 1 002 的四个子集进行实验,实验数据集中包含对象数量分别为 200、400、600、800,检出的人脸图片数量分别为 4 166,8 313,12 502,16 670。分别对四个实验数据集使用三折交叉验证。不同数据集下的分类准确率与 Yaw 方向角度的平均绝对误差分别如图 6 和图 7 所示。

实验结果表明,随着实验数据集中对象数量的增加,数据集中包含的具有代表性的信息越多,在相同的网络结构下,测试图片的分类准确率逐渐增加,平均绝对误差逐渐减小。当样本对象数量超过 600 后,实验

结果有所下降。

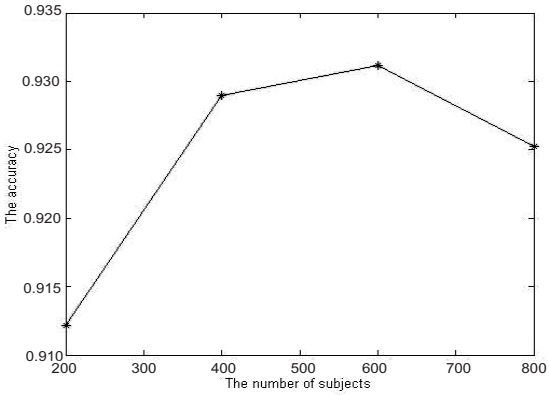

图 6　不同规模数据集下的分类准确率

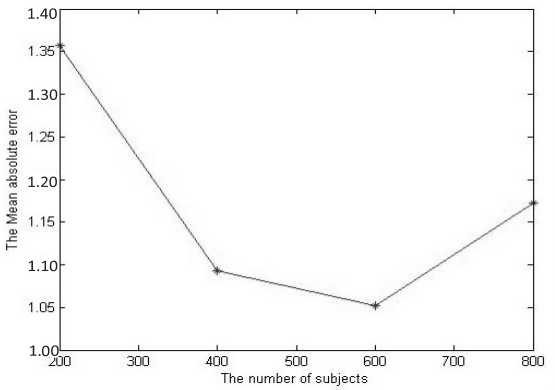

图 7　不同规模下的平均绝对误差

4　结束语

文中提出了基于深度学习的头部姿态估计方法。通过深度网络提取特征并对其进行分类预测，在 CAS-PEAL 数据集上显著降低了姿态估计的误差，取得了较好的实验效果。但是该方法的泛化能力强弱依赖于训练数据的多样性与网络结构的复杂度。由于训练图片来自 CAS-PEAL 数据集，若测试图片来自其他数据集，则测试效果不理想。

下一步的工作主要是融合多个数据集训练更加复杂的网络结构，在保证较低的姿态角度误差的前提下，增大网络结构的泛化能力。

参考文献：

[1] Kuchinsky A, Pering C, Creech M L, et al. FotoFile: a consumer multimedia organization and retrieval system[C]//Proceedings of the SIGCHI conference on human factors in computing systems. New York: ACM, 1999: 496-503.

[2] Wang Jiangang, Eric S. EM enhancement of 3D head pose estimated by point at infinity[J]. Image and Vision Computing, 2007, 25(12): 1864-1874.

[3] Ebisawa Y. Head pose detection with one camera based on pupil and nostril detection technique[C]//Proceedings of the IEEE international conference on virtual environments, human-computer interfaces and measurement systems. [s. l.]: IEEE, 2008: 172-177.

[4] Kong S G, Mbouna R O. Head pose estimation from a 2d face image using 3D face morphing with depth parameters[J]. IEEE Transactions on Image Processing, 2015, 24(6): 1801-1808.

[5] Haj M A, Gonzalez J, Davis L S. On partial least squares in head pose estimation: how to simultaneously deal with misalignment[C]//Proceedings of IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2012: 2602-2609.

[6] Foytik J, Asari V K. A two-layer framework for piecewise linear manifold-based head pose estimation[J]. International Journal of Computer Vision, 2013, 101(2): 270-287.

[7] Lu J, Tan Y P. Ordinary preserving manifold analysis for human age and head pose estimation[J]. IEEE Transactions on Human-Machine Systems, 2013, 43(2): 249-258.

[8] Fanelli G, Dantone M, Gall J, et al. Random forests for real time 3D face analysis[J]. International Journal of Computer Vision, 2013, 101(3): 437-458.

[9] Ma B, Chai X, Wang T. A novel feature descriptor based on biologically inspired feature for head pose estimation[J]. Neurocomputing, 2013, 115: 1-10.

[10] Geng X, Xia Y. Head pose estimation based on multivariate label distribution[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2014: 1837-1842.

[11] Le Cun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[12] Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2013: 3476-3483.

[13] Yan S, Shan S, Chen X, et al. Matrix-Structural Learning (MSL) of cascaded classifier from enormous training set[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2007.

[14] Ma B, Huang R, Qin L. VoD: a novel image representation for head yaw estimation[J]. Neurocomputing, 2015, 148: 455-466.