基于外观的视线估计方法综述

徐立明

北京科技大学自动化学院北京 100083

【摘　要】本文针对基于外观的视线跟踪方法进行了综述，对此类技术的发展、相关研究工作和研究现状进行了阐述。然后就基于外观的视线估计方法中存在的几个关键性问题进行了探讨，并将目前的研究方向分成了三大类，分别是：减少环境依赖、减少训练样本以及解决视线估计中的头动问题，并从这三个角度出发，对现有的视线跟踪方法进行了对比分析,此外,还介绍了视线跟踪技术的分类以及在多个领域的应用，最后对基于外观的视线跟踪方法的发展趋势和研究热点进行了总结与展望。

【关键词】视线估计；基于外观；映射模型；深度学习；

中图分类号：TP391　文献标识码：A

引言：视线估计技术目前被广泛应用于研究个体视线方向、视线或关注点（POR），是确定人的视觉注意力的技术。近年来，视线跟踪系统的发展和应用愈发引起关注[1]，许多学者对人眼的状态识别和跟踪进行了探索,此方向也是近年来机器视觉的一个研究热点，其研究成果覆盖了许多的领域，如人机交互[2]，虚拟现实[3]，视频眼科，残疾人辅助[4]和人类行为研究等。

视线估计方法通常分为基于模型或基于外观的方法。基于模型的方法使用眼睛和面部的几何模型来估计注视方向，通常依赖于瞳孔检测以及光源在角膜上的反射形成的普尔钦斑[5]，通过计算眼睛运动特征建立从几何模型的瞳孔中心到凝视校准点的映射。这些方法可以达到很高的精度，但是需要复杂的设备以及精确的校准过程[6]，并不具有普遍适用性。基于外观的方法不是利用几何映射关系，而是将裁剪后的眼睛图像对应于高维空间中的点，进而学习从给定特征空间中的点到屏幕坐标的映射关系。因此，可以通过隐式建模的方法获取输入眼睛图像的视线估计方向。一般的基于外观的视线估计方法基于使用整个眼睛图像的强度特征（直方图信息或梯度信息）作为输入数据，输出为用户视觉关注点或视线方向。

基于外观的视线估计与几个应用领域相关，包括基于凝视的人机交互和视觉行为分析[7]，在计算机视觉领域中的研究中显得尤为重要。文献[8]中提出了基于学习的方法学习来自大量用户的与头部姿势无关的训练数据。这种方法的提出加速了基于外观的视线估计技术在手持和便携式设备中的应用进程，例如手机和笔记本电脑以及交互式显示器等。就目前看来尽管基于外观的视线估计方法在日常环境中表现良好，但仍然存在着诸多问题，比如很多已经提出的方法中没有在不同的数据集上进行评估，这会带来重大数据集偏差的风险,这一问题在目标识别[9]和目标显著性检测[10]领域也属于关键问题。

1. 研究现状

在基于外观的视线估计中往往会存在如下问题，首先最基本的问题是仅使用少量训练样本进行准确的凝视估计以便于校准；其次是当头部的运动较难预测时提出算法来抵消掉由于自由头部自由运动带来的误差；最后的是实际的视线估计系统对于光照等条件的变化要具有鲁棒性。

1.1 减少环境依赖

为了减少方法对人员和场景的依赖的限制，文献[11]提供了一种解决的新方法,引入仅仅来自于“未经校准的注视模式”的独立于人和场景眼睛图像。适用于通过单眼相机在未指定的场景中捕获的常见眼睛图像。与之相反，以前的方法获得的数据需要在特定场景中通过某些硬件设施捕获。通过非线性降维和像素运动分析恢复人员的注视模式，无论个体和所在场景如何变化，都仅从眼睛图像中恢复未校准的注视模式，这种方法仅仅依靠尺度变换和平移就可以得到注视点位置，传统方式一般依靠足够的训练数据来计算人和场景特定的非线性凝视映射，显然该方法比传统的校准方式简单高效得多。而且所提出的方法可以通过简单的校准将未校准的凝视模式与真实世界的凝视位置（例如，屏幕坐标）对齐。与传统方法相比，校准灵活，任务不中断；可以比传统校准快几个数量级，并且标定过程需要最少甚至零用户参与。上述优点令此方法从根本上与其他方法不同，同时也更具有实际意义。例如，所提出的未校准的凝视模式满足许多应用中的要求，例如在认知分析和视频监视领域应用传统的校准方法就十分困难。尽管如此，此方法传统方法大致相似，头部较大的姿势变化会增加误差。

通常情况下，所提出的基于外观的视线估计算法都是在实验室条件下得到的数据集上进行验证或者并未在多个数据集中进行评估，这无疑是基于外观的视线估计技术迈向实用化的一大障碍。文献[12]研究了在野外的视线估计技术，不再将实验环境限制在实验室内，提出的方法在一定程度上克服了光照等条件的影响。

所提供 MPIIGaze 数据集包含实验人员从 15 名参与者中采集到的 213,659 张图片，时间跨度超过三个月。因此，该数据集比现有数据集在外观和照明方面更具可变性。提出了一种使用多模卷积神经网络进行基于外观的视线估计的方法，首先采用 SURF 级联方法[13]进行人脸检测来定位从校准的单目 RGB 相机获得的输入图像中的地标。 然后，拟合通用 3D 面部形状模型来估计检测到的面部的 3D 姿势，并应用文献[14]中提出的空间归一化技术来将头部姿势和眼睛图像裁剪和扭曲到标准化训练空间。CNN 用于学习从头部姿势和眼睛图像到摄像机坐标系中的视线方向的映射。该方法在 MPIIGaze，EYEDIAP[15] 和 UT Multiview[14]三个数据集上进行了广泛的评估。

文献[16]提出了一个基于 16 层 VGG 深度卷积神经网络[17]的深度外观视线估计方法–GazeNet,在同样的数据集上 GazeNet 的技术水平提高了 22%（对于最具挑战性的跨数据集评估，从平均误差 13.9 度到 10.8 度）。与文献[16]相比，本文做出的改变有：

（1）扩展注释 37677 个图像，包括六个面部标志（四个眼角和两个角落）和瞳孔中心；

（2）更新网络架构到 16 层 VGGNet；

（3）在合成数据训练时进行新的交叉数据集评估；

（4）评估与领域无关的视线估计条件，特别是视线范围的差异、照明条件和个人形象；

（5）评估图像分辨率，是否使用双眼，以及头部姿势和瞳孔中心信息等对最终的视线估计效果的影响。

1.2 减少训练样本

文献[18]中提出基于一种 l^1–optimization framework 的自适应线性回归方法(ALR)来解决上述问题。该方法利用稀疏和低维训练样本来预测输入眼图像中的注视点位置。从这个意义上讲，可以显着减少所需训练样本的数量。 而且，在相同的优化框架下，提出了一种精确的子像素对齐方法，该方法推广了 ALR 的基本优化框架。对齐和视线估计同时进行以处理由于轻微头部运动引起的问题，对

于图像分辨率的变化带来的问题也有显著改善。在同一优化框架内提出了眨眼检测方法，它可以检测眨眼，同时不会受到注视方向变化的干扰。它适用于每个图像，无需来自相邻帧的信息。该文献中提出的基于外观的视线估计方法可以应用于在只需要简单校准的系统中，并且对多种影响因素具有鲁棒性。此外，所提出的方法能够快速适应其他类似的问题。但是，方法仍然存在局限性，首先，此方法只能很好地处理轻微的头部运动，对于较大的头部自由移动则无能为力，而文献[19]提出结合头部姿势和眼睛中心位置来计算凝视方向。此方法是基于特征的，在理论上与基于外观的方法不同。但是利用基于特征的方法处理较大的头部运动和基于外观的方法进行视线估计不失为一个很好的未来的发展方向。其次，虽然减少了训练样本的数量，但仍然是在屏幕坐标系中使用 2D 回归获取注视点位置，因此需要相机相对于屏幕的固定关系，因此仅限于特定的设备配置，即无法直接推广到其他设备，不具有实用性。

眼睛的图像是计算机视觉中的关键问题，得益于与计算机图形学的并行进步的眼区建模技术，无约束的视线估计技术得到了很好的发展。针对这些问题的大规模监督方法需要耗时的数据收集和手动注释，在实际应用中，这显然是不合理的。而文献[20]使用计算机图形技术来构建头部扫描的动态眼睛区域几何模型集合，应用眼睛模型合成大量高度逼真的眼睛区域图像，这些图像适用于各种头部姿势，凝视方向和照明条件，从而显著减少数据收集和注释工作。考虑到面部运动和眼球旋转所经历的动态形状变化以及眼球本身的复杂材料结构，眼睛区域特别难以精确建模。为了解决这个问题，提出了一种使用动态集合大规模渲染逼真的眼睛图像和眼区模型的新方法。本文进行了交叉数据集实验来检验所提出的方法的效果。使用与 UT 数据集相同的相机设置合成了训练图像。然后训练卷积神经网络（CNN）模型并评估它们在 MPIIGaze 数据集上的表现。经过实验验证，CNN 在本文的 SynthesEyes 数据集上训练的模型的性能（μ=13.91◦）与作为模型训练的 UT 数据集上性能（μ=13.55◦）相差不多。因此证实了本文合成图像的方法是有效的。

虽然可穿戴或短距离（小于 60cm）远程眼动仪的视线估计技术或多或少已经很成熟了，但由于高分辨率眼图不可用，对于中远距离情景下的视线估计仍然是一个具有挑战性的问题。因为在许多实际场景中，例如人机交互、第一人称视觉和数字标牌系统等，只有低分辨率的眼睛图像可用。因此解决低分辨率眼睛视线估计的问题是必不可少的。文献[14]介绍了一种用于精确视线估计的合成学习方法，且与个体和个体头部姿势无关。与现有的基于外观的方法中需要个体独立的训练数据不同，文中使用大量的跨个体训练数据来训练 3D 视线估计网络。收集较大且完全校准的多视图注视数据集并执行 3D 重建，以生成眼睛图像的密集训练数据。通过使用合成数据集来学习随机回归森林，实验结果显示该方法在使用低分辨率眼睛图像的方法中较为优秀。

1.3 解决头动问题

考虑到头部的自由移动进行精准的视线估计在诸如驾驶员监视，广告分析和监视的场景中具有广泛的应用，可靠且低成本的单眼解决方案对于这些领域的普遍使用至关重要。为了解决如文献[11]和文献[15]中提到的的头动问题，文献[21]提出了一种结合深度特征提取和特征森林回归的新型视线估计方法，并致力于开发一种在自然光下利用一个单独的网络摄像头的凝视跟踪系统。与一般的基于外观的视线估计方法略有不同，所提出的方法从 CNN 中提取深度特征以预测眼睛注视方向。其中 CNN 通过多尺度卷积将眼睛图像进行分类，并将最后隐藏层作为深度特征汇集。深度特征的引入在标记图像和图像分类方面等方面有着重要意义。提取了深层特征后，使用回归森林来寻找特征空间和凝视方向之间的直接相关性。在森林节点分裂期间，通过集群最小化误差平方和二进制分类对节点进行分割，以便在深度特征空间中实现更好的分区。微调聚类算法和 SVM 分类的应用对凝视回归中子节点误差的降低具有积极影响，因为它对最终聚类得到的视线估计结果有均衡效果。所提出的方法在自然光和自由头部运动的条件下在眼睛图像数据集上表现良好，对实验中存在的某些遮挡的影响也具有一定的鲁棒性。

注意到视线向量是头部姿势和眼球运动根据一定几何关系的组合，文献[22]提出了一种新颖的注视变换层来连接单独的头部姿势和眼球运动模型。 所提出的方法避免了头部姿势–注视方向相关性的过度拟合，并且能够普遍应用于各种训练集。为了加强对网络训练的监督，提出了一个两步训练策略，首先使用粗糙标签训练子任务，然后用准确的注视标签进行联合训练。为了减少主体和环境变化对最后结果的影响，本文建立了一个大型数据集，其中包括头部姿势和眼球运动，包含 200 个主体在不同的照明条件下的数据。经过实验验证，该方法实现了较高的注视跟踪精度，使用在单个 CPU 上以 1000fps 运行的小型网络（不包括面部对齐时间）达到 5.6° 跨主体预测误差，对于深度训练网络这一误差缩小至 4.3° 。

.2.总结

基于外观的视线估计方法主要依靠得到的眼睛区域的图像计算视线方向，最初的基于 2d 回归的算法十分不准确，在此基础上，实验人员做出的提高有：采集多个主体在多种头部姿态下的数据，利用深度学习网络得到不局限于主体与环境的视线估计方法；为了针对大规模监督学习的方法需要耗时的数据收集和手动注释，应用已扫描得到的眼睛模型合成大量高度逼真的眼睛区域图像，利用合成图像进行训练，得到网络应用于视线估计；利用深度特征提取和随机森林等方法解决头部自由移动的问题。

.参考文献

[1] Sun L, Liu Z, Sun M T. Real time gaze estimation with a consumer depth camera[J]. Information Sciences, 2015, 320(C):346–360.

[2]张建兴. 基于注意力的目标识别算法及在移动机器人的应用研究[D]. 重庆大学，2013.

[3] Reale M J, Liu P, Yin L, et al. Art critic: Multisignal vision and speech interaction system in a gaming context.[J]. IEEE Transactions on Cybernetics, 2013, 43(6):1546–1559.

[4] Päivi Majaranta. Twenty years of eye typing:systems and design issues[C]// Symposium on Eye Tracking Research & Applications. 2002:15–22.

[5] 胡艳红, 魏江, 梅少辉. 基于瞳孔角膜反射技术的视线估计方法[J]. 计算机工程与应用, 2018, v.54；No.909(14):12–15+23.

[6] Ki J, Kwon Y M, Sohn K. 3D Gaze Tracking and Analysis for Attentive Human Computer Interaction[C]// Frontiers in the Convergence of Bioscience & Information Technologies. IEEE, 2007:617–621.

[7]Morimoto C H , Mimica M R M . Eye gaze tracking techniques for interactive applications[J]. Computer Vision and Image Understanding, 2005, 98(1):4–24.

[8]Funes Mora K A , Odobez J M . Person Independent 3D Gaze Estimation From Remote RGB–D Cameras[C]// IEEE International Conference on Image Processing. IEEE, 2014.

[9]A. Torralba and A. A. Efros– Unbiased look at dataset bias.In Proc. CVPR, pages 1521–1528. IEEE, 2011.

[10]Li Y , Hou X , Koch C , et al. The Secrets of Salient Object Segmentation[J]. 2014.

[11] Lu F, Chen X, Sato Y. Appearance–Based Gaze Estimation via Uncalibrated Gaze Pattern Recovery.[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2017, PP(99):1–1.

[12]Zhang X , Sugano Y , Fritz M , et al. Appearance–Based Gaze

（下转第 201 页）

（上接第 195 页）

Estimation in the Wild[J]. 2015.

[13] Li J , Zhang Y . Learning SURF Cascade for Fast and Accurate Object Detection[C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2013.

[14] Sugano Y, Matsushita Y, Sato Y. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2014.

[15]Mora K A F, Monay F, Odobez J M. EYEDIAP:a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras[C]// Symposium on Eye Tracking Research & Applications. 2014.

[16]Zhang X, Sugano Y, Fritz M, et al. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation[J]. IEEE Trans Pattern Anal Mach Intell, 2017, PP(99):1-1.

[17]F. Lu, Y. Sugano, T. Okabe, Y. Sato, Adaptive linear regression for appearance-based gaze estimation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (10)(2014) 2033 - 2046.

[18] Valenti R, Sebe N, Gevers T. Combining Head Pose and Eye Location Information for Gaze Estimation[J]. IEEE Transactions on Image Processing, 2012, 21(2):802-815.

[19]Wood E , Baltrusaitis T , Zhang X , et al. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation[J]. 2015.

[20]K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition// in Proc. Int. Conf. Learning Representations, 2015.

[21] Wang Y, Shen T, Yuan G, et al. Appearance-based gaze estimation using deep features and random forest regression[J]. Knowledge-Based Systems, 2016, 110(C):293-301.

[22] Deng H, Zhu W. Monocular Free-Head 3D Gaze Tracking with Deep Learning and Geometry Constraints[C]// IEEE International Conference on Computer Vision. 2017.