1. **Introduction**

视线估计是一种旨在理解用户意图和兴趣的技术[1]，已经成为人机交互、心理学、计算机视觉、医学诊断等诸多领域的重要工具。人们使用了不用的方法来处理视线估计，然而，有些方法在实际情况下是不准确的，而有些方法需要显式的用户校准，这样就会很麻烦。其中一些方法存在图像质量差和光照条件变化的问题。最近深度学习的成功和普及极大地提高了视线估计的性能。大规模数据集的可用性进一步提高了及与深度学习的方法的性能。视线估计技术关注的是图像数据和注视方向或者注视点之间的关系[2]。根据从单个或者多个摄像头采集的眼睛区域的图像数据（如瞳孔或者角膜反射）来估计注视方向[3]。目前，存在不同类型的商业和非商业的视线估计解决方案，包括基于模型的方法和基于外观的方法。本文将详细的介绍基于外观的方法，同时简略介绍基于模型的方法，这是因为基于模型的方法需要外部设备对人进行固定，或者需要多种设备对视线估计的结果进行精确计算，这在很多现实场景中是很难实现的。

1. **Methods**

视线估计方法通常可以分为基于模型的（model-based）和基于外观的（apperence-based）[2]。

1. 基于模型的视线估计（MODEL-BASED GAZE ESTIMATION）

基于模型的方法是使用一个眼睛的几何模型，利用人眼成像的原理，通过相机获取人眼眼球的特征，随后根据获取到的特征建立与注视点的映射关系，从而实现实现估计。基于模型的方法可以进一步分为基于角膜反射的方法和基于形状（shape-based）的方法。角膜反射法通过外部红外光源反射到眼睛的角膜来检测到眼睛的特征。

(2)基于外观的视线估计（APPEARANCE-BASED GAZE ESTIMATION）

基于外观的视线估计方法不依赖于眼睛特征点的检测，而是直接从眼睛光学外观来执行视线估计的[2]。通常需要一台摄像机来捕捉眼睛的图像，然后用它来简历视线估计模型，将捕捉到的图像外观映射到特定的注视方向或者注视点。这些模型不需要手工设计特征进行训练，它们可以从数据中隐式的提取图像特征。最早的基于外观的视线估计工作是将眼睛图像作为神经网络的输入，用于100个分类任务，每个类别代表一个注视方向[4]。Tan等人[5]是第一个将基于外观的视线估计作为论文工作的，该论文将校准样本视为流形中的点，并试图寻找权重来插值测试样本。本文将基于外观的视线估计的基本形式定义为特征提取和从特征到注视目标的回归映射。接下来的工作分别使用高斯过程回归[6]，神经网络[7]，和自适应线性[8]等方法，沿用原始像素作为他们的特征输入。这些早期的基于外观的视线估计方法都假定头部姿势是固定的，以此来减少外观的变化，避免复杂的讨论。

随后的研究开始允许自由的3D头部移动。Lu等人[8]将原始的头部自由运动问题分解为了两个子问题，分别是初始固定头部问题和随后的补偿来矫正初始估计误差，将由头部姿势引起的注视偏差作为最终视线估计结果的补偿。

尽管基于外观的视线估计技术取得了很大的进展，但这些方法比传统方法需要更多的训练数据来学习头部自由运动引起的眼睛外观的显著变化。因此，这些研究主要针对于特定领域或者个人进行评估。视线估计中的一个开放性研究挑战是学习没有用户、环境和相机限制的视线估计器。这些估计器可以在不需要额外输入的情况下推广到任何人。Schneider等人[9]使用了哥伦比亚数据集（Columbia dataset）[10]对流形对齐方法进行了评估，发现在6种回归方法中，使用流形对齐方法可以提高视线估计性能。Zhang等人[11]为了避免摄像头的偏移和人头部姿态对视线估计的影响，首先对眼部数据进行人脸对齐和3D头部估计、眼部图片归一化，将单眼图像作为输入，通过LeNet-5网络对图像特征进行提取，并求出的二维头部偏转信息合并到视线估计网络的全连接层，依据混合特征进行视线估计。该论文同时提出了MPIIGaze数据集，该数据集包括21万张人脸图片，涵盖15位参与者和不同光照条件。Krafka等人[12]通过IOS应用收集了GazeCapture数据集，包含1450多人、240多万张数据，并提出iTracker网络，将人左右眼部、脸部图片和脸部网格图片分别输入到模型当中，最终预测相对于摄像头的凝视位置信息，在iPhone上实现了1.71厘米的估计误差，在iPad上达到了2.53厘米的估计误差。

Deng和Zhu[13]提出了一个视线变换模型，将头部姿势和眼球运动模型连接起来，并收集了一个大的数据集，包括200个志愿者和不同光照条件。在另一个研究中，Kim等人[14]提出了另一个基于回归的视线估计模型，他们受到了iTracker网络的启发，设计了名为Gazelle的模型，该模型需要五个输入，分别是：右眼、左眼、脸、脸网格和梯度直方图(HOG,Histogram of Gradients)，更好的捕捉用于视线估计的特征。在上述模型中，人脸图像主要提供头部姿态信息，而脸部网格则提供眼部位置信息，存在一定的信息冗余，因此，He等人[15]对上述模型的输入进行了简化，将脸部图片和脸部网格图片这两个输入替换为四个眼角的坐标位置信息，因为眼角坐标不止提供了眼睛的位置信息，还暗含了头部姿态信息，即眼角间距越小，头部姿态越大。实验结果表明，该简化后的模型在iPhone上的误差达到了1.78cm，和原始模型相差不多，同时改模型的计算效率很高，处理速度可以达到10ms/帧。

随后，Zhang等人[16]尝试将AlexNet修改为16层的VGG网络，该研究将平均误差从13.9度提升至10.8度。2017年，Zhang等人[17]认为基于学习的方法可以利用其它面部区域的附加特征，不同于之前手动提取眼部和脸部图像信息，他们尝试只将人脸图像作为输入，利用CNN对人脸图像进行编码，同时提出了一种空间权重机制，将人脸不同区域的信息有效地编码到CNN架构中，该机制在卷积层的特征图上学习空间权重，来灵活的抑制或者增强不同区域的信息。经过广泛的评估，该方法在MPIIGaze和EYEDIAP上分别有14.3%和27.7%的改进。因为数据集收集繁琐和标注困难等问题，Wood等人[18]开发了一个动态眼部区域模型，可以生成不同姿态的带有标记的可训练图像，从而大大减少了数据的收集和标注工作。在此基础上，Wood等人[19]提出了UnityEyes，一种可以快速合成大量可变眼域图像的新方法，该方法结合了一个人眼区域生成三维模型和一个实时绘制框架，并将该框架免费在线提供，方便研究使用。最新的方法是通过对抗训练模型[20]或者生成模型[21]来生成仿真图像。

2018年，Liu等人[22]提出一种新的用于视线估计的图像差分方法，直接训练一个卷积神经网络来预测同一个被摄体的两个眼部输入图像之间的注视差异，随后使用一组标准参考图片来对视线预测进行修正，Liu等认为使用深度神经网络构建的回归器，基于外观的方法精度被限制在5-6°左右，不同用户之间具有不同的差异，因为具有相同外观但内部结构不同的两个眼睛图像可以对应不同的注视方向，解决这个问题的一个简单的方法就是学习特定于人的模型，但是训练特定的模型需要大量的个人数据，这在实际中是不现实的。作者提出了两种修正的方法，第一种是将[11]提到的视线估计模型作为基础模型，通过数据分析发现，预测的视线向量和真实值之间可以拟合为线性关系，因此在基础模型输出时加上一个线性映射来修正预测值，并通过最小均方误差(LMSE)来作为损失函数进行训练。第二种方法差分神经网络预测视线向量差值，利用孪生网络，将同一个眼部不同图片选择两张作为输入，使网络学习两张图片的差别，并输出两张图片视线估计向量的差值，随后在测试中分别将需要预测的眼部图片和一组标定的参考图片作为模型输入，将模型输出的视线向量差值加上参考视线向量并求取平均值作为眼部图片的视线预测结果。Liu等利用少量标定图像建立了特定对象的模型来改进基于外观的凝视估计，保证预测结果更加稳定。

2020年，Seonwook等人[23]提出了一种新的EVE数据集（End-to-end Video-based Eye-tracking）和相应的方法。该数据集收集了54名参与者，超过1200万帧，总计约105小时的视频数据。模型分为两个部分，分别是EyeNet和GazeRefineNet，EyeNet负责对每帧进行视线估计和瞳孔大小估计，GazeRefineNet通过合并屏幕内容、时间信息和EyeNet得到的初始注视点信息来优化最后的注视点预测结果。最终方法使得注视点估计正确率提高了28%，角度误差减小到了2.49度。2020年，Yu等人[24]尝试使用无监督的方式进行视线估计，网络需要输入样本，目标样本以及输入样本和目标样本之间的注视角度（包括垂直方向角度和水平方向角度）的差值。网络通过解析输入样本和差值对输入图像进行像素维度的重定向，从而得到接近于目标图像的输出。通过缩小输出图像和目标图像之间的差异，迫使网络学习到视线估计相关的表征，这样就实现了无监督学习方法的视线估计。

1. **Conclusion**

从结果来看，基于模型的视线估计技术比基于外观的技术会有更小的预测误差。大多数基于外观的或者基于CNN的技术都会产生大于0.5°的预测误差。主要原因之一是缺乏分布均衡且可变的大规模数据集，这些数据集中的图像是从多个设备捕获的，但是来自每个设备的数据的可变性和可用作训练的数据数量并没有得到很好地平衡。为了获得非常好的性能，来自每个设备的训练数据也应该具有良好的外观、头部姿态、亮度和方位的可变性，未来的工作可以进行收集平衡的大数据集。

其次是目前不进行特定用户修订的方法的错误角度大多徘徊在4-5°，数据集标定也是造成该现象的一个原因，因为收集数据集后需要人为标定，会造成数据标定的错误问题，这样也会限制模型的学习和正确率的提高。

大多数基于CNN的技术都是针对无约束环境设计的，在这种环境中，捕捉到的图像可能会受到不同头部姿势、外观、照明、面部遮挡等影响。

目前提出的一些有监督的视线估计方法都是在相对较小的数据集上训练的，一般来说，要建立有效且稳健的深度学习模型，需要有良好平衡的和大规模的数据集。

对于数据集收集困难，标注麻烦的问题，从结果上看，基于合成数据的方法和基于迁移学习的方法，目前还难以接近使用真实数据训练的模型。从这个角度看，无监督或自监督学习可能会是视线估计领域的发展方向。保证在仅有少数个性化用户样本时，可能对模型进行微调，以提高模型的正确率。

1. **Reference**
2. Tsukada A, Shino M, Devyver M, et al. Illumination-free gaze estimation method for first-person vision wearable device[C]//2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, 2011: 2084-2091.
3. Hansen D W, Ji Q. In the eye of the beholder: A survey of models for eyes and gaze[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32(3): 478-500.
4. Chennamma H R, Yuan X. A survey on eye-gaze tracking techniques[J]. arXiv preprint arXiv:1312.6410, 2013.
5. Baluja S, Pomerleau D. Non-intrusive gaze tracking using artificial neural networks[C]//Advances in Neural Information Processing Systems. 1994: 753-760.
6. Tan K H, Kriegman D J, Ahuja N. Appearance-based eye gaze estimation[C]//Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings. IEEE, 2002: 191-195.
7. Williams O, Blake A, Cipolla R. Sparse and Semi-supervised Visual Mapping with the S^ 3GP[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 1: 230-237.
8. Liang K, Chahir Y, Molina M, et al. Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression[C]//Proceedings of the 2013 Conference on Eye Tracking South Africa. 2013: 17-23.
9. Lu F, Okabe T, Sugano Y, et al. Learning gaze biases with head motion for head pose-free gaze estimation[J]. Image and Vision Computing, 2014, 32(3): 169-179.
10. Schneider T, Schauerte B, Stiefelhagen R. Manifold alignment for person independent appearance-based gaze estimation[C]//2014 22nd international conference on pattern recognition. IEEE, 2014: 1167-1172.
11. Smith B A, Yin Q, Feiner S K, et al. Gaze locking: passive eye contact detection for human-object interaction[C]//Proceedings of the 26th annual ACM symposium on User interface software and technology. 2013: 271-280.
12. Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4511-4520.
13. Krafka K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2176-2184.
14. Zhu W, Deng H. Monocular free-head 3d gaze tracking with deep learning and geometry constraints[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3143-3152.
15. Kim M, Wang O, Ng N. Convolutional neural network architectures for gaze estimation on mobile devices[J]. Standford Univ., Standford, CA, USA, Tech. Rep, 2016.
16. He J, Pham K, Valliappan N, et al. On-device few-shot personalization for real-time gaze estimation[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019: 0-0.
17. Zhang X, Sugano Y, Fritz M, et al. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1): 162-175.
18. Zhang X, Sugano Y, Fritz M, et al. It's written all over your face: Full-face appearance-based gaze estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 51-60.
19. Wood E, Baltrusaitis T, Zhang X, et al. Rendering of eyes for eye-shape registration and gaze estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3756-3764.
20. Wood E, Baltrušaitis T, Morency L P, et al. Learning an appearance-based gaze estimator from one million synthesised images[C]//Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. 2016: 131-138.
21. Shrivastava A, Pfister T, Tuzel O, et al. Learning from simulated and unsupervised images through adversarial training[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2107-2116.
22. Wang K, Zhao R, Ji Q. A hierarchical generative model for eye image synthesis and eye gaze estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 440-448.
23. Liu G, Yu Y, Mora K A F, et al. A Differential Approach for Gaze Estimation with Calibration[C]//BMVC. 2018, 2: 6.
24. Park S, Aksan E, Zhang X, et al. Towards End-to-end Video-based Eye-Tracking[J]. arXiv preprint arXiv:2007.13120, 2020.
25. Yu Y, Odobez J M. Unsupervised Representation Learning for Gaze Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7314-7324.