3D Gaze Direction Estimation from Surveillance Camera Videos Based on Full-Body Posture and Temporal Rotation Modeling

氏　　名　　上村 敬志

研 究 室 名　　サイバネティクス・リアリティ工学研究室

主指導教員名　　清川 清

内 容 梗 概

This dissertation presents 3D gaze direction estimation for people in surveillance-camera videos. Unlike conventional gaze estimation, surveillance-camera images are often captured from elevated, distant viewpoints, making facial and eye regions low-resolution, occluded, or invisible. We focus on methods that use full-body posture and temporal information rather than detailed facial appearance.

This research was developed in three stages. First, a geometry-based method was investigated, in which 3D human posture was estimated from surveillance-camera images and gaze direction was approximated from head direction using 3D keypoints such as the ears and nose. This approach showed that gaze direction can be estimated without eye-region information, but its accuracy was limited by pose estimation errors and the simplified assumption that head direction represents gaze.

Second, an end-to-end deep learning method was developed to estimate 3D gaze direction from a single surveillance-camera image. A 3D pose estimation framework on deep learning was used to extract full-body information, and the model directly inferred gaze direction from image input. This method improved applicability to surveillance-camera environments because it did not require explicit head or body annotations.

Finally, we extended the single-frame method to a multi-frame framework that models temporal changes in gaze direction. The proposed framework estimates a reference gaze direction from the central frame, predicts inter-frame rotations on SO(3), represents these rotations using the Matrix-Fisher distribution, and integrates the gaze sequence with an LSTM. This probabilistic temporal modeling improves geometric consistency and reduces error accumulation across frames, achieving state-of-the-art accuracy on a surveillance-oriented dataset using only image inputs.

Overall, this research establishes a progression from geometric posture-based gaze estimation to single-frame deep learning and probabilistic temporal modeling. The results demonstrate that robust 3D gaze estimation from distant full-body surveillance imagery is feasible and that temporal rotation modeling improves accuracy under low-resolution, occluded, and non-frontal conditions.