機械学習による視線推定と
その実世界応用

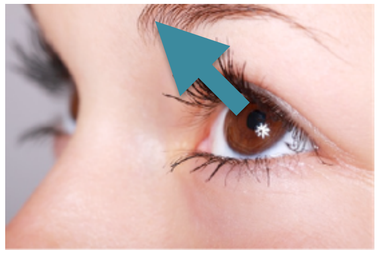
菅野裕介

大阪大学大学院 情報科学研究科

大阪大学
OSAKA UNIVERSITY

視線推定・アイトラッキング

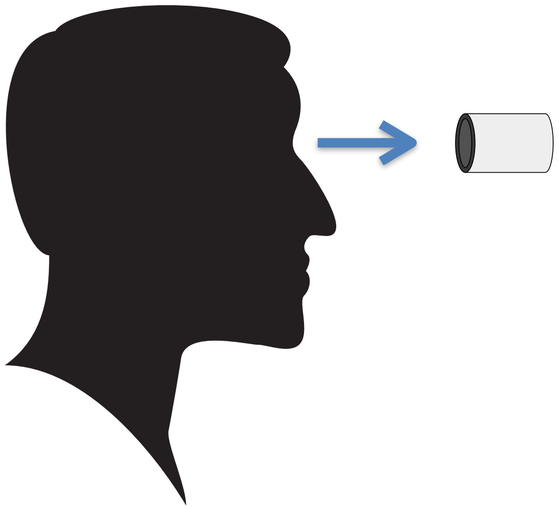
- 人物がどこを見ているかを計測・推定する
  - Attentive User Interfaces
  - 広告・マーケティング調査
  - 人間の内部状態を推定するための特徴として



- 従来の手法は専用のハードウェアを必要とする

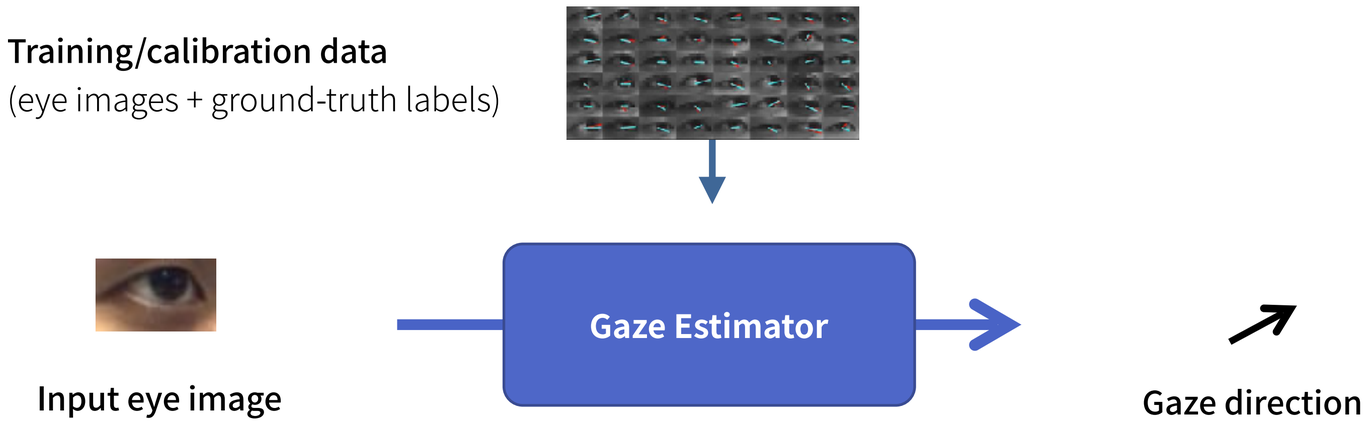
- 日常生活空間でも簡単に使える視線推定を実現するには？

カメラベースの視線推定



- 通常のカメラから、画像中に写っている人物の (3次元) 視線方向を推定する
  - ウェブカメラ、ウェアラブルカメラ、…
- 従来手法では実現できないアプリケーションが多数
  - HRI、一人称視点映像解析、公衆空間での注意推定

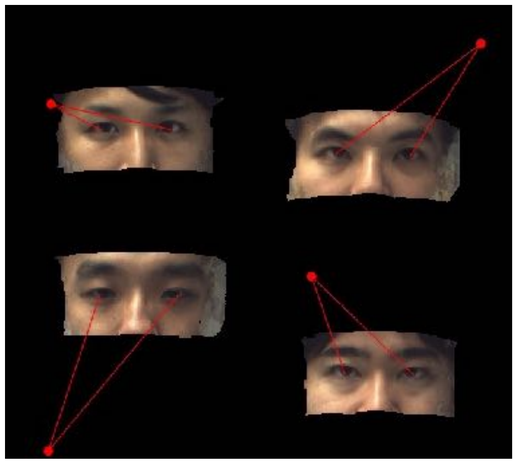
アピアランスベース視線推定



- アプローチ自体は古くから存在 [e.g., Tan et al., WACV'02]
  - 初期の手法は人物ごとの学習データを想定していた
- 人物非依存の推定器を学習することはできる？
  - 以下の要素をカバーする学習データセットが必要
    - 人物ごとに異なる目の形状、カメラに対する顔向きの違い、照明条件・撮影環境の違い、…

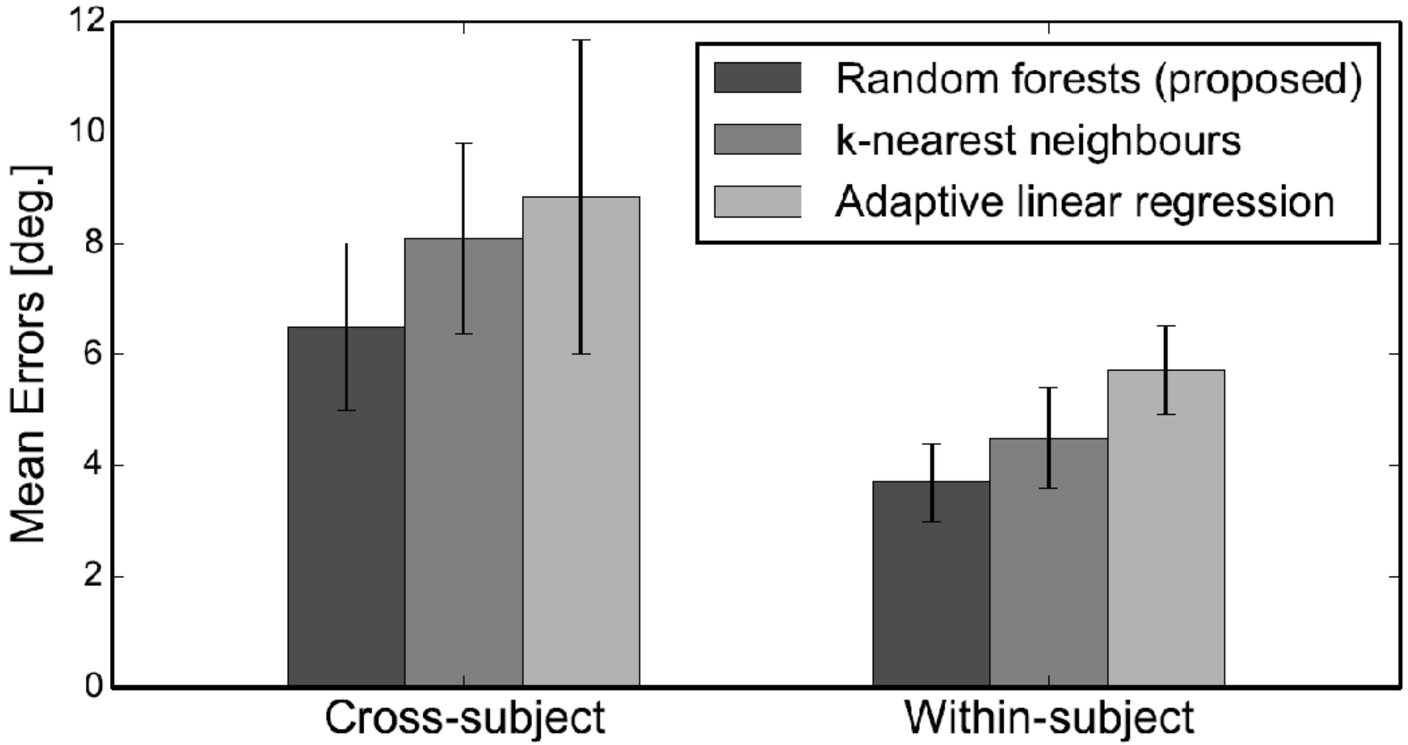
Multi-view gaze dataset [Sugano et al., CVPR'14]

- 8台のカメラを同期して撮影した視線真値付きの目画像データセット
  - 50人 × 160 (16 × 10) 視線方向
- 3D reconstruction can be used to synthesize training data
  - 50 × 160 × 144姿勢 = 1,152,000枚の目画像
- 人物非依存の視線推定関数を学習することが可能



データセット上での推定性能

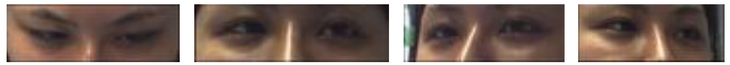
- 平均誤差: 6.5 度
  - 3-fold cross-person validation: 生成データで学習、実データで推定
  - 個人ごとのキャリブレーションを行っていないことを考えると十分な性能？



アピアランスベース視線推定 “in the wild”

- 実応用上はこれよりも遥かに多様な入力画像を想定する必要がある

Cross-domain training



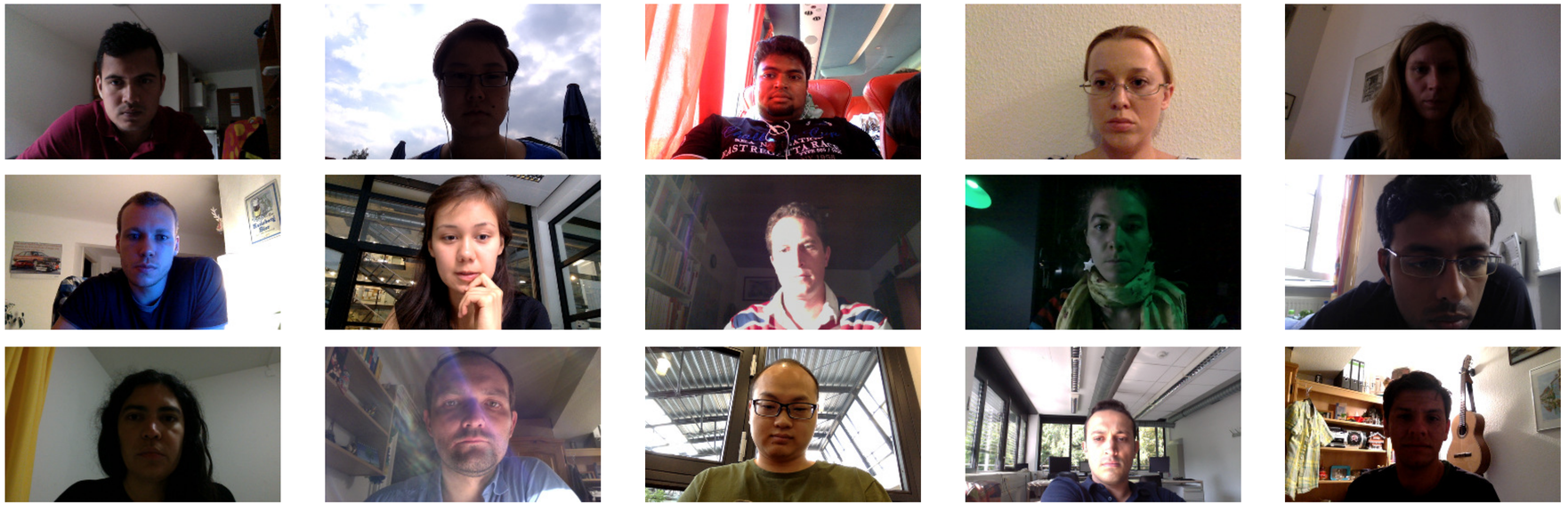
- 外部データセットで学習
- 未知の環境に対する頑健性

Within-domain training

- 対象環境で獲得したデータで学習
- 性能上限の評価

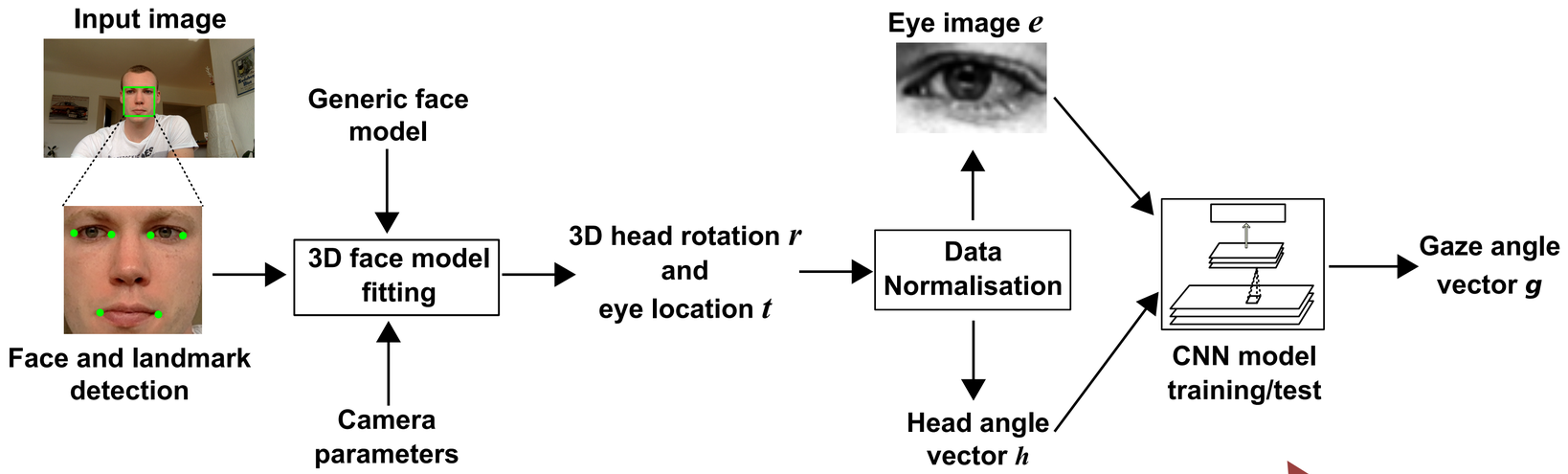
Daily life environment

MPIIGaze dataset [Zhang et al., CVPR'15, TPAMI (accepted)]

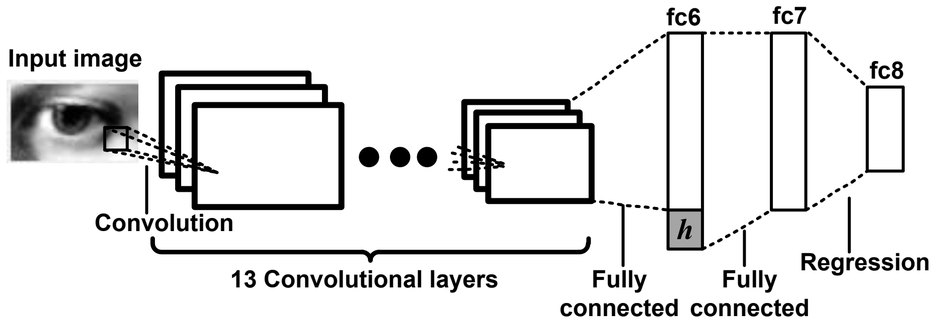


- より実環境に近い多様な照明条件下で撮影したデータセット
  - 15人、～3ヶ月、計213,659画像
  - ラップトップPCにインストールしたソフトウェアで視線真値付きの画像を定期的に撮影
  - 日常生活内で定期的に撮影が行われることで撮影環境の多様性を担保

畳み込みニューラルネットワークによる視線推定



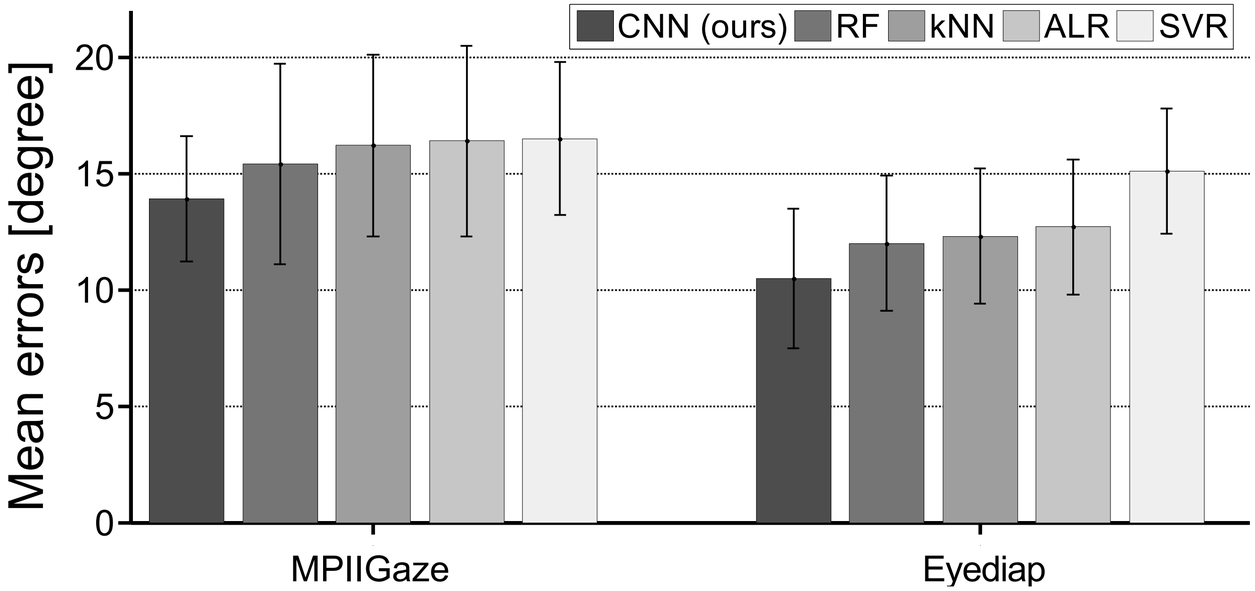
顔と顔器官の検出
+ 3次元顔形状モデルのフィッティング



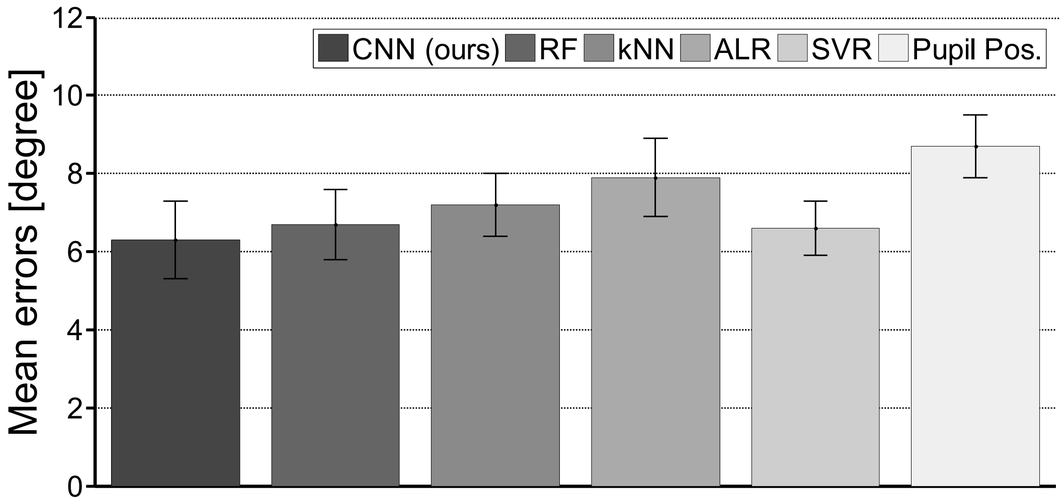
CNNアーキテクチャ
(目画像入力、全結合層で頭部姿勢ベクトルを結合)

性能の比較

Cross-domain evaluation (Training on UT Multiview)



Within-domain evaluation (on MPIIGaze)



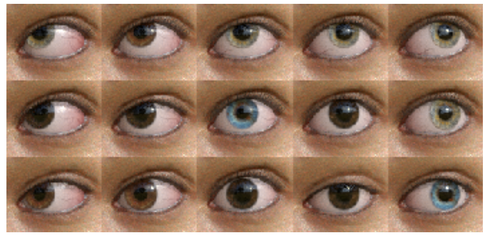
- 学習データセットの違いによる性能の違いが顕著
  - 学習データが対象ドメインから得られている場合、同等の性能が得られる
  - 異なるデータセットでの学習条件の場合にCNNベースのアプローチが有利

学習データの獲得

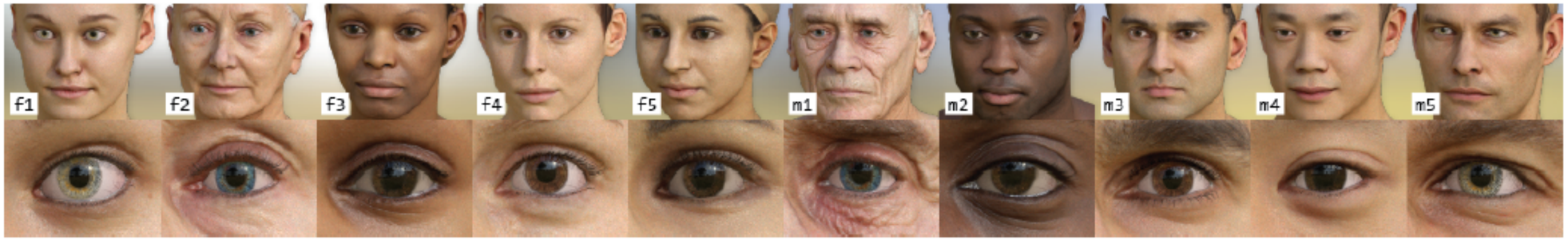
- 照明条件の変化が重要な要素の一つ

  - 対象となる環境から学習データを獲得できれば解決するが、現実的には難しい
  - 新しい環境に対応できる推定器を事前学習するには？

- カメラ/頭部姿勢だけではなく、学習データの照明環境もコントロールする
  - 3DCGモデルを使った学習データ生成

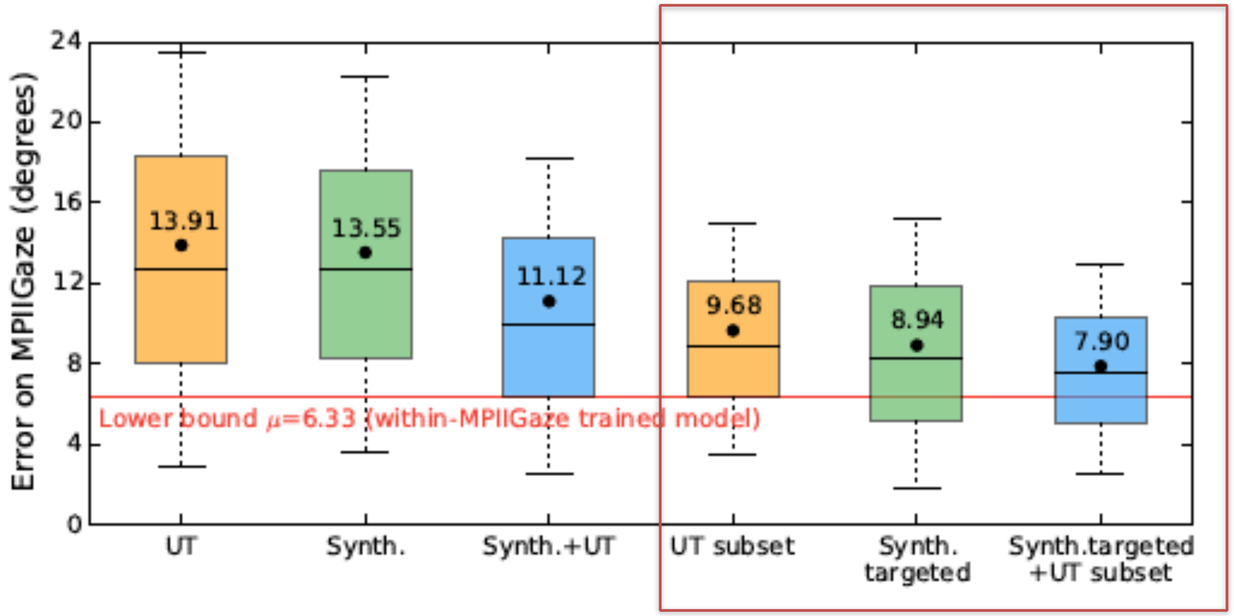


SynthesEyes Dataset [Wood et al., ICCV'15]



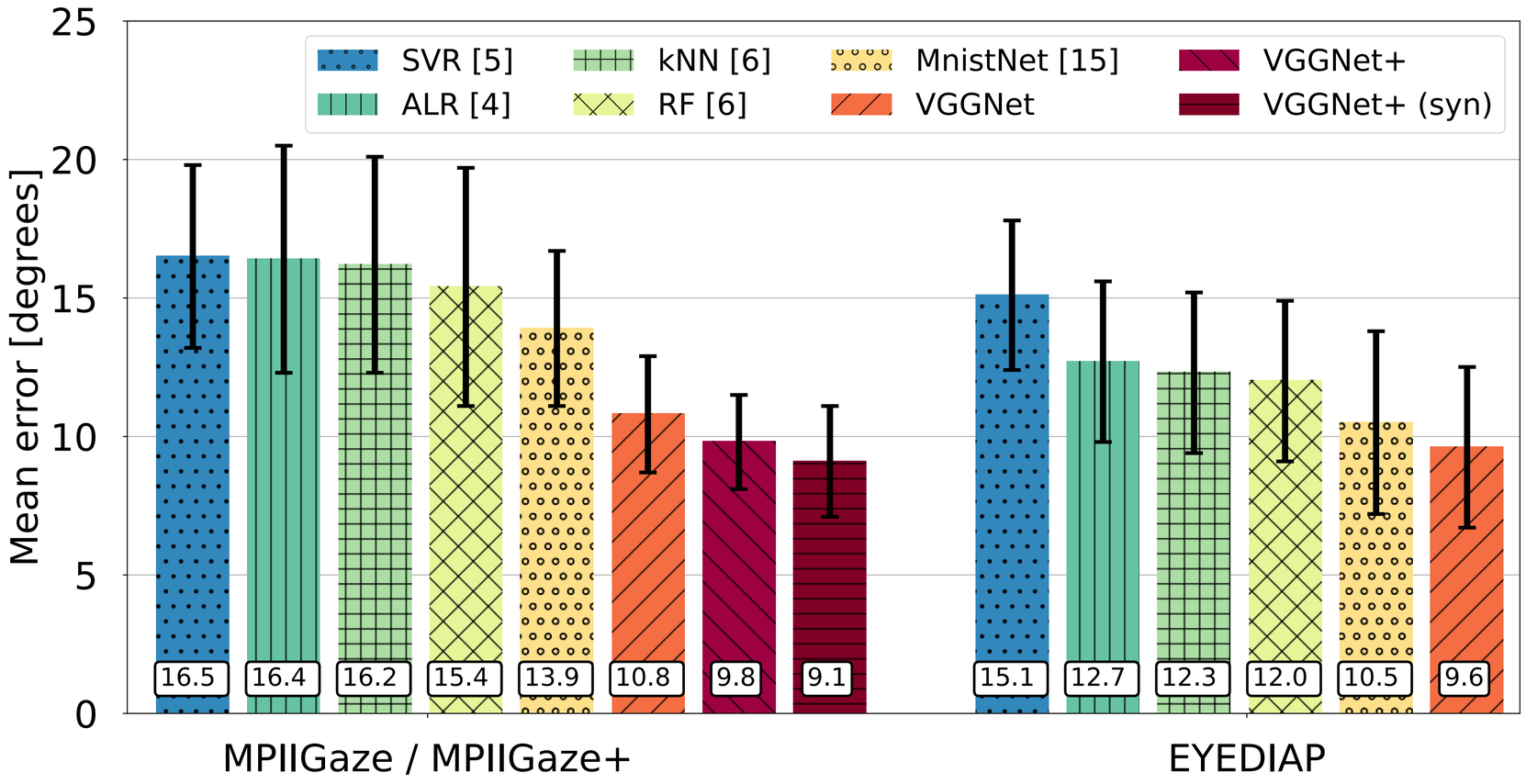
- 顔の3次元形状・テクスチャモデルと眼球モデルの組み合わせ
  - イメージベースライティングによる照明条件コントロール
- Multiviewデータセットと比較して、
  - 照明も変化させながら、正確な視線方向と共に目画像を生成することができる
  - 顔・目領域形状の多様性は制限される (形状変化のモデルは無い)

性能比較



- MPIIGazeデータセット上での推定性能 (UT Multiview vs. SynthesEyes)
  - SynthesEyes (3DCGベース) による学習結果はUT Multiview (画像ベース) とほぼ同等
- CG画像で事前学習したモデルを実画像でFine Tuningしたモデルが最も高い精度を実現
- 照明条件以上に、視線方向の範囲を制限することの効果が大きい

性能比較

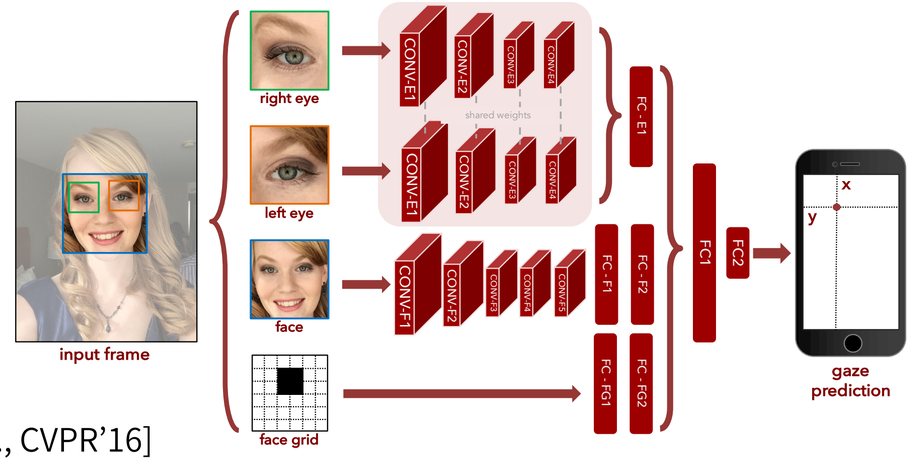


- より深いネットワーク構造にすることで最終的な推定誤差は〜9度

顔画像入力からの視線推定 [Zhang et al., CVPRW'17]

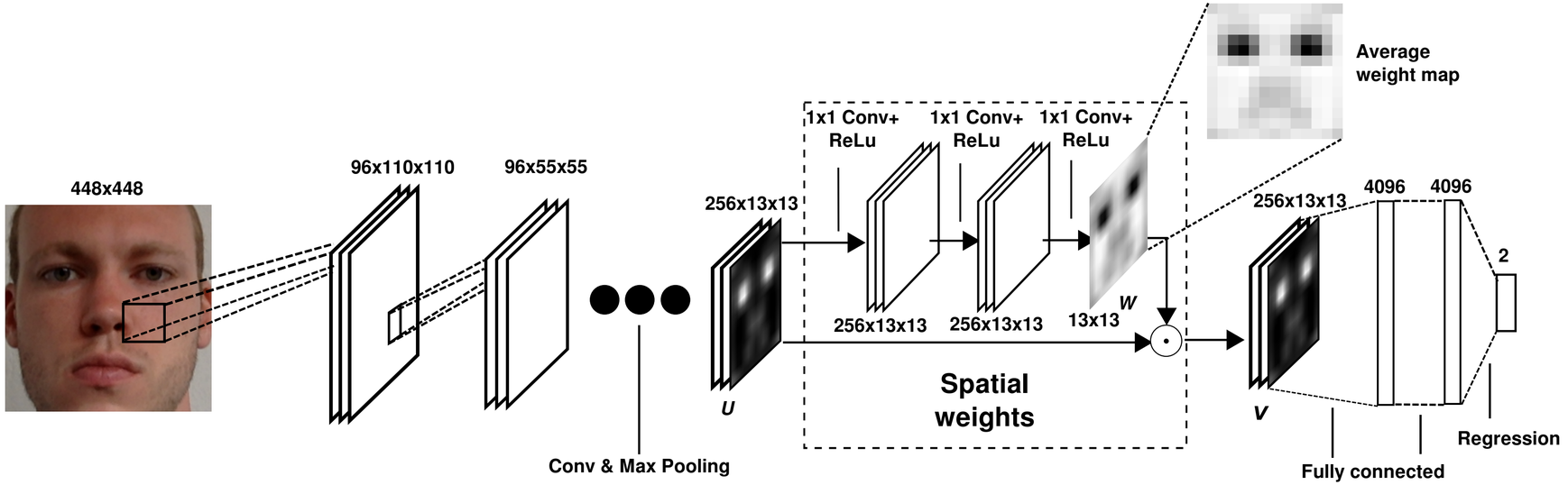
- これまでのモデルは片目毎に視線方向を推定
  - 物理的には正しいが、多くの場合視線方向が両目で一致しない

- 顔画像全体をそのまま入力にすれば良いのでは？
  - パーツ毎の入力を結合するアーキテクチャによる性能向上事例
  - 最適な組み合わせを手動で設計している



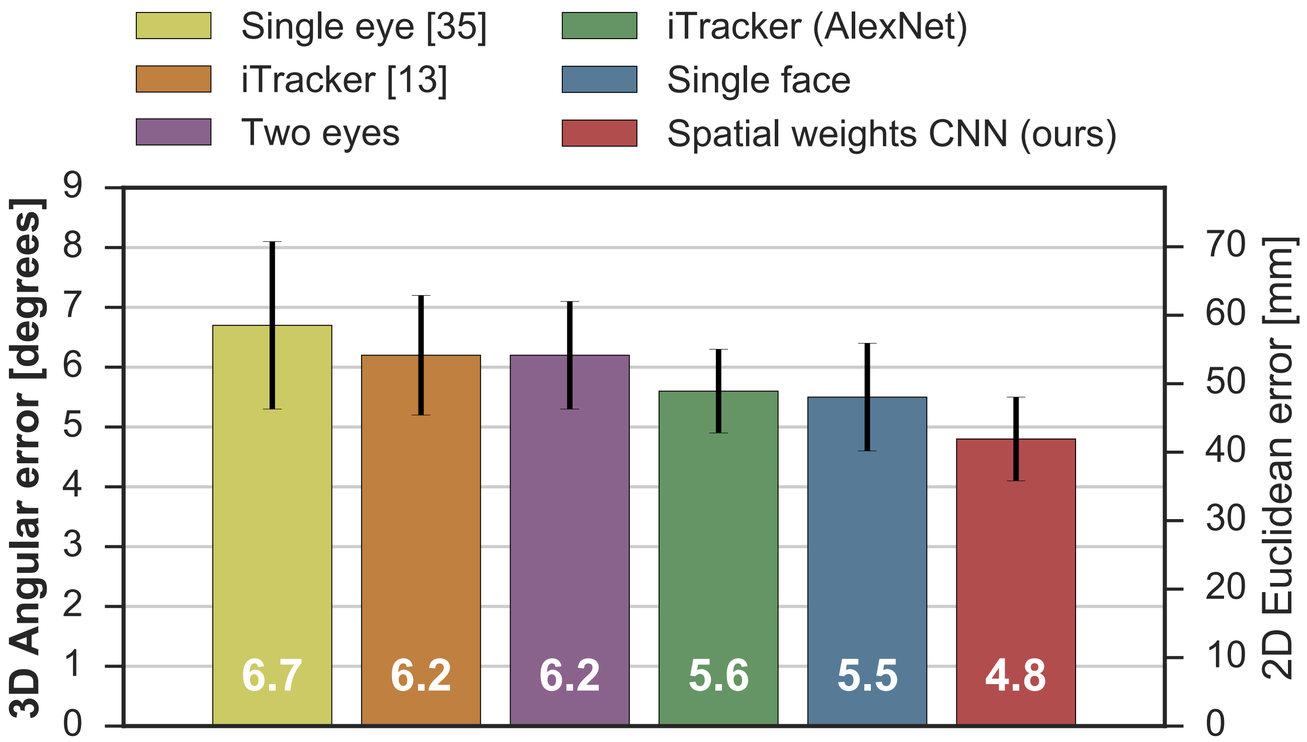
[Khosla et al., CVPR'16]

視線推定CNNのための空間的な重みの学習



- 目以外の領域の重要性は入力画像に依存する
  - 目画像は常に重要
  - 例えば、極端な頭部姿勢を持つ入力画像では目以外の領域の重要度が増す
- 視線推定CNNに、空間的な重み推定層を追加して学習する
  - 顔領域の重要な特徴を強調する

推定性能



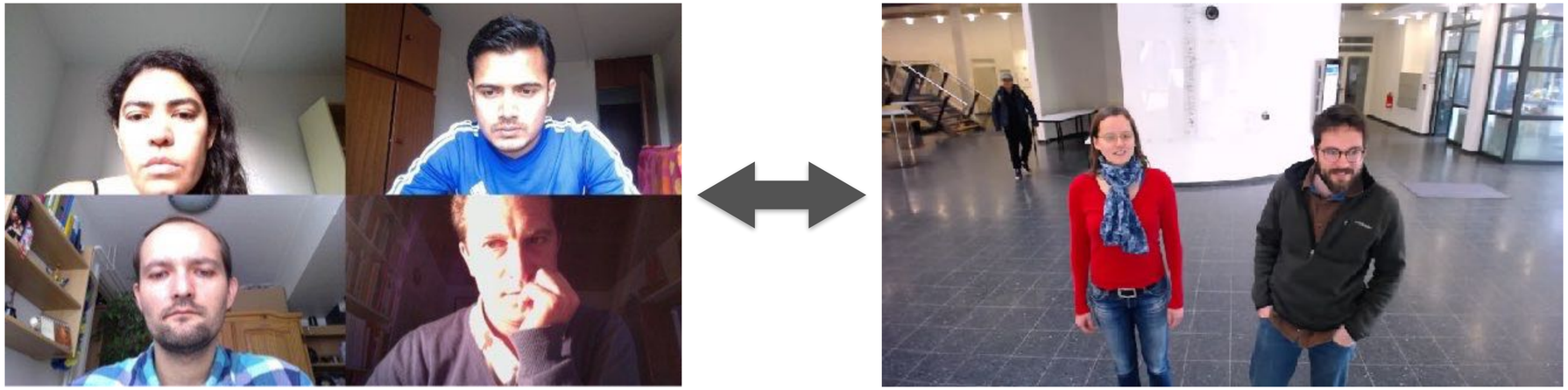
- 平均誤差4.8度 (MPIIGazeデータセット内学習・評価の場合)
  - 手作業で設計された顔画像統合モデルより高い性能

アピアランスベース視線推定の実世界応用

- 画像ベースの認識結果を実際のアプリケーションに使うことはできる？

技術的課題

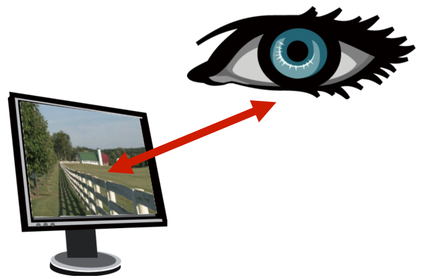
- 学習環境とテスト・応用環境の違い
  - ドメイン間の違いに起因する問題は完全には解決できていない

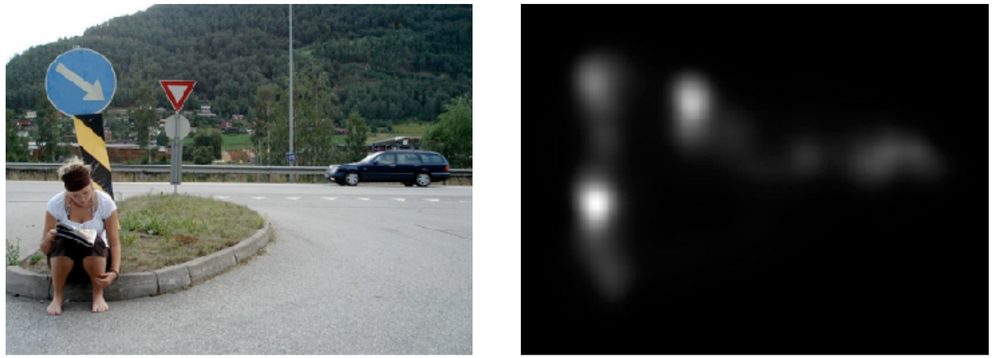


- 理想的な条件でも精度は限られる
  - 従来の視線推定応用が想定している推定誤差は～1度

HCIの観点から適応的なシステムを設計するアプローチ

- 対象の環境・ユーザから学習データを獲得できるよう、システムの設計自体を工夫する
  - システムが意図している「視線方向真値」にユーザの注意を誘導する
  - ユーザの行動や人間的特性からその人がどこを見ているかを予測する

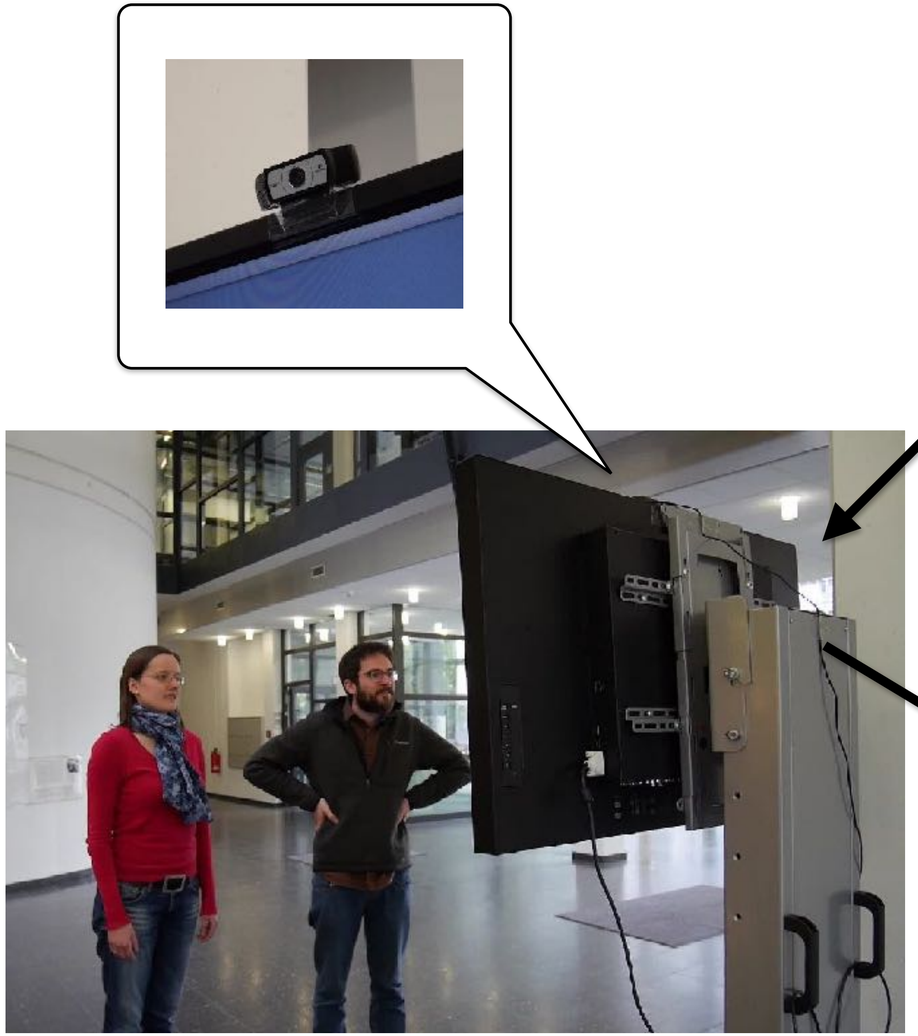




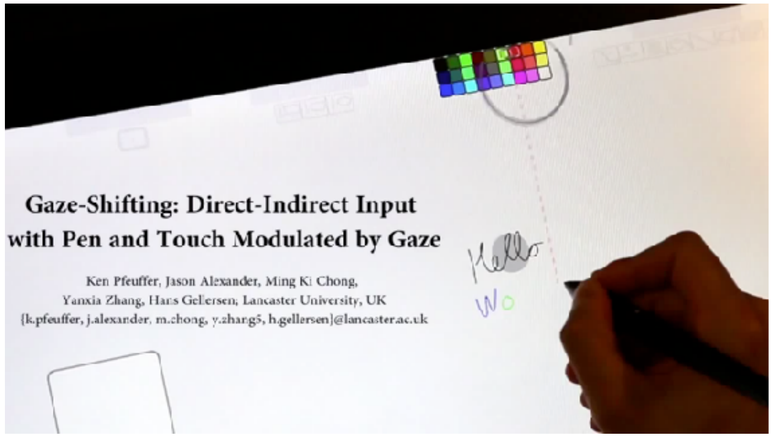
視覚的顕著生マップ
(ボトムアップな注意推定)
[Sugano et al., CVPR'10, TPAMI'13, UIST'15]

インタラクション情報
(マウスクリックなど)
[Sugano et al., ECCV'10, THMS'15]

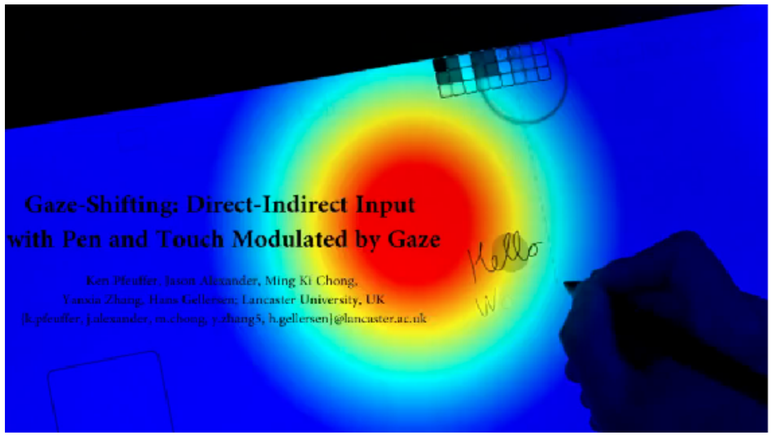
パブリックディスプレイでの視線推定 [Sugano et al., UIST'16 (best paper honorable mention)]



Stimuli

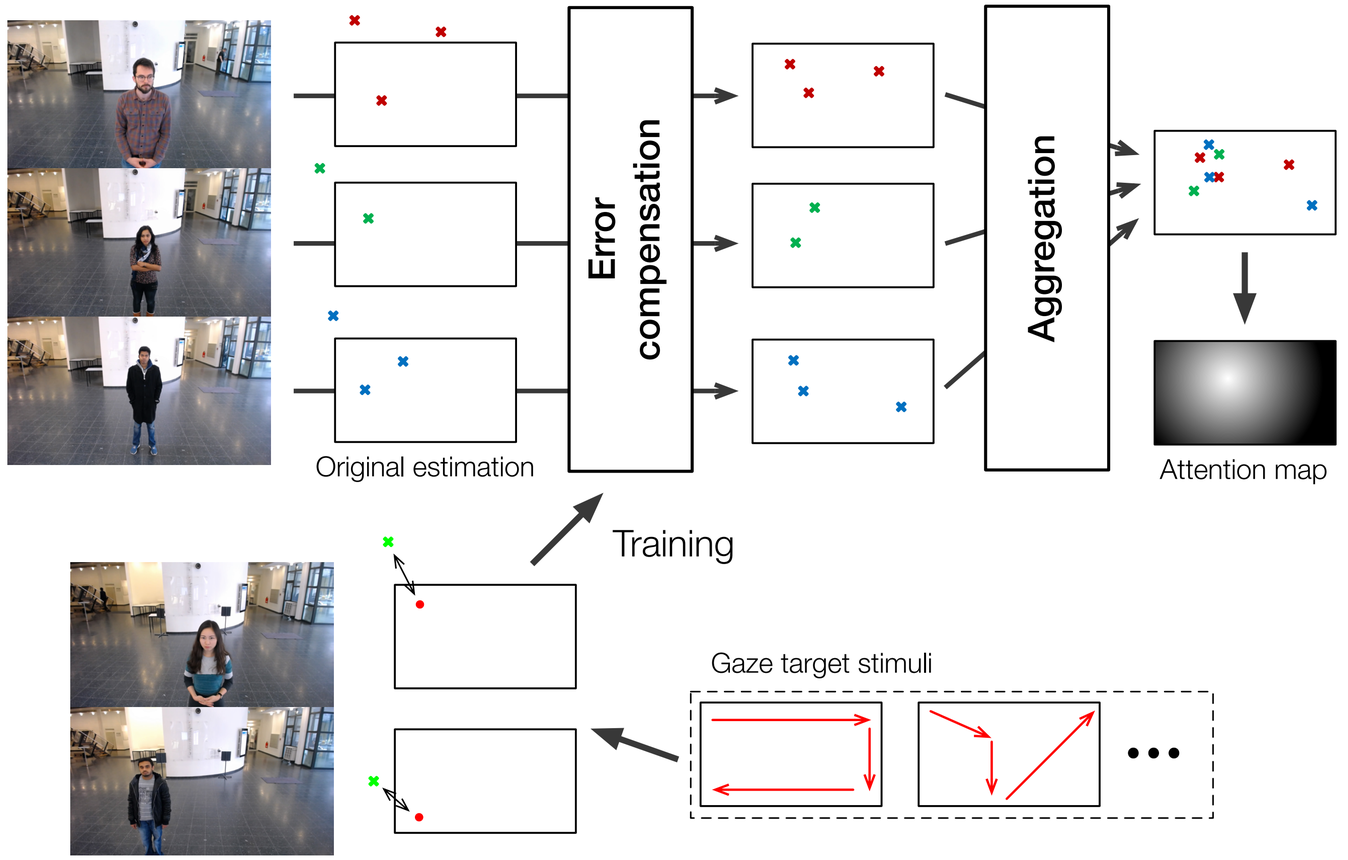


Attention maps



基本的なアイデア

- **環境**に起因する誤差に対処するために、補正関数を追加で学習する
- **個人**に起因する誤差に対処するために、複数人の推定結果を統合する



誤差補正関数の学習データ

1. 運用前に「キャリブレーション」を行う

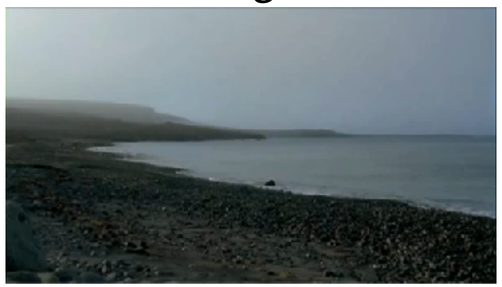
- 何人かテストユーザを募って、通常の視線推定学習と同様にターゲットを見てもらえば良い

誤差補正関数の学習データ

1. 運用前に「キャリブレーション」を行う
   - 何人かテストユーザを募って、通常の視線推定学習と同様にターゲットを見てもらえば良い
2. 運用中にユーザの注視位置が予測できるコンテンツを挿入する

Stationary

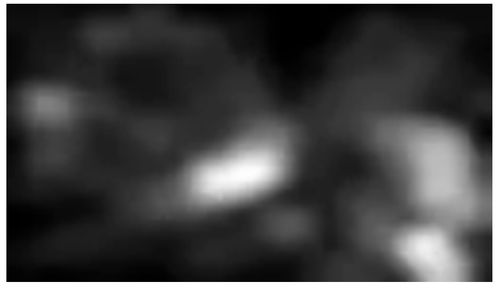
Edge



Gaze template/Saliency

Pursuit

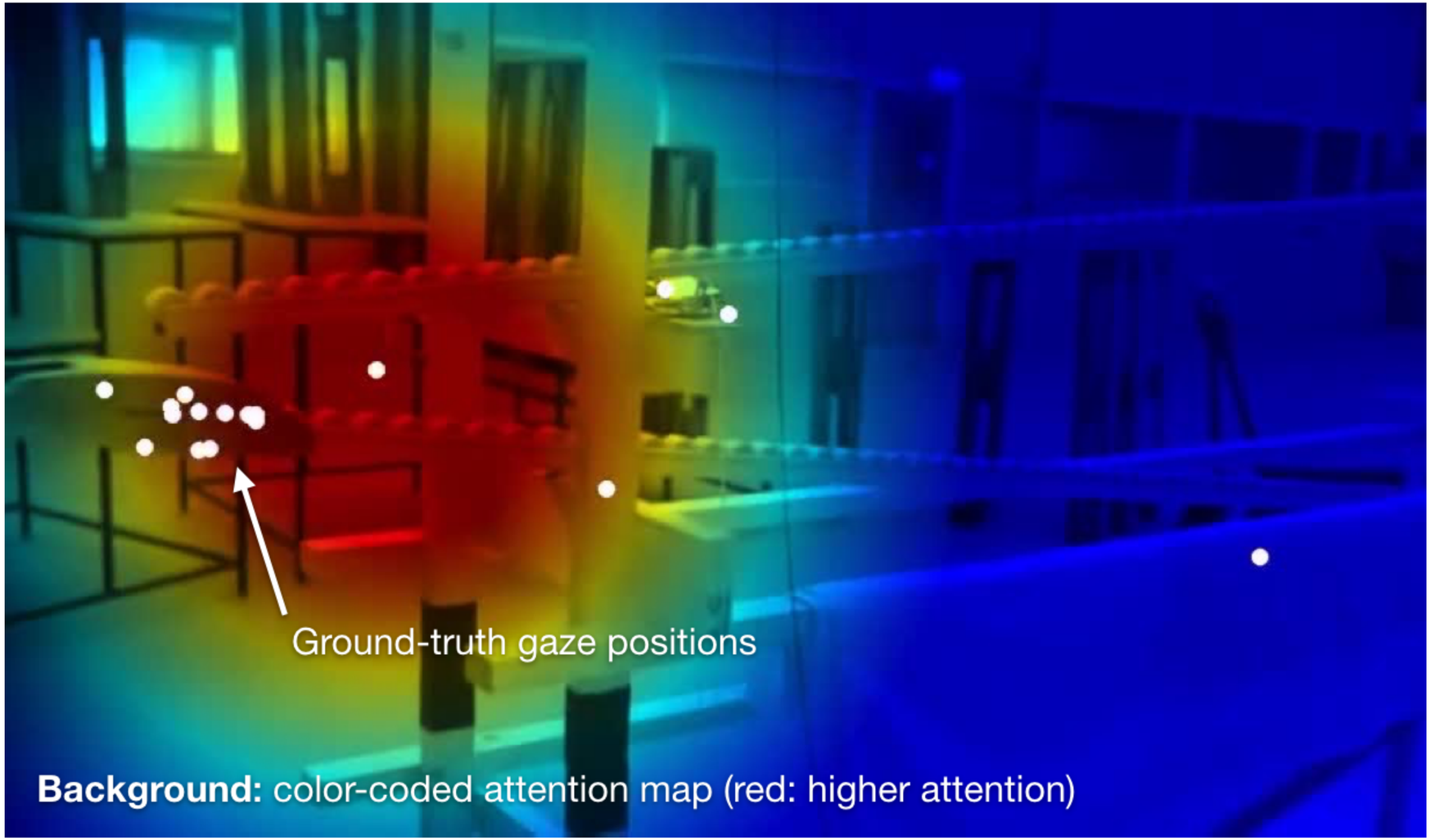
Text



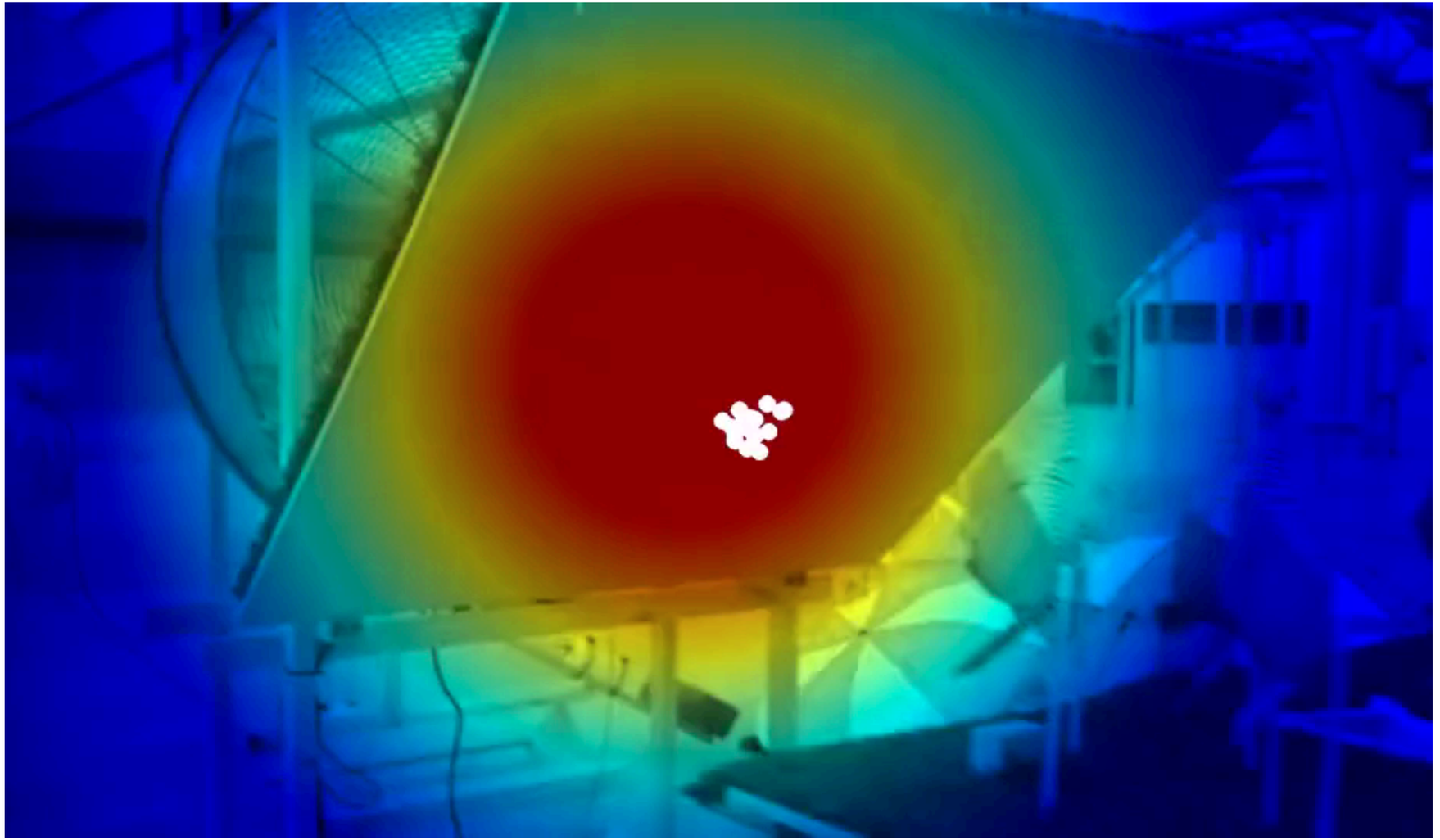
運用・評価実験

- パブリックスペースに設置したシステムで不特定多数の注視を計測
  - 複数のビデオをループで流し続けている大型ディスプレイにウェブカメラを設置
  - **12時間×13日**、1ビデオあたり平均44人の顔を検出・統合
  - 商用アイトラッカーによる推定結果と比較

Attention Map推定結果の例



Attention Map推定結果の例

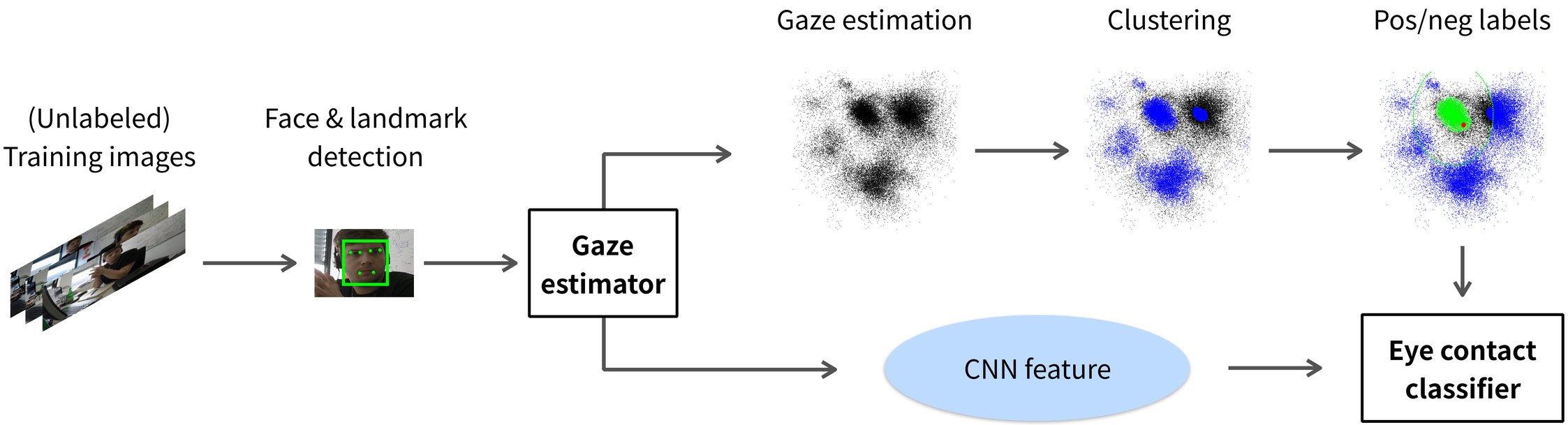


アイコンタクト検出 [Zhang et al., UIST'17 (best paper honorable mention)]

- ユーザが対象物体を見ているかどうかを識別する
  - HRIなど、アプリケーションによってはこれで十分なケースもある
- アイコンタクト検出は視線推定より容易なタスク？
  - 連続値の推定→二値分類
  - アイコンタクト検出器を事前学習するアプローチ [e.g., Smith et al., UIST'15]
- 識別境界は常に対象物体に依存する

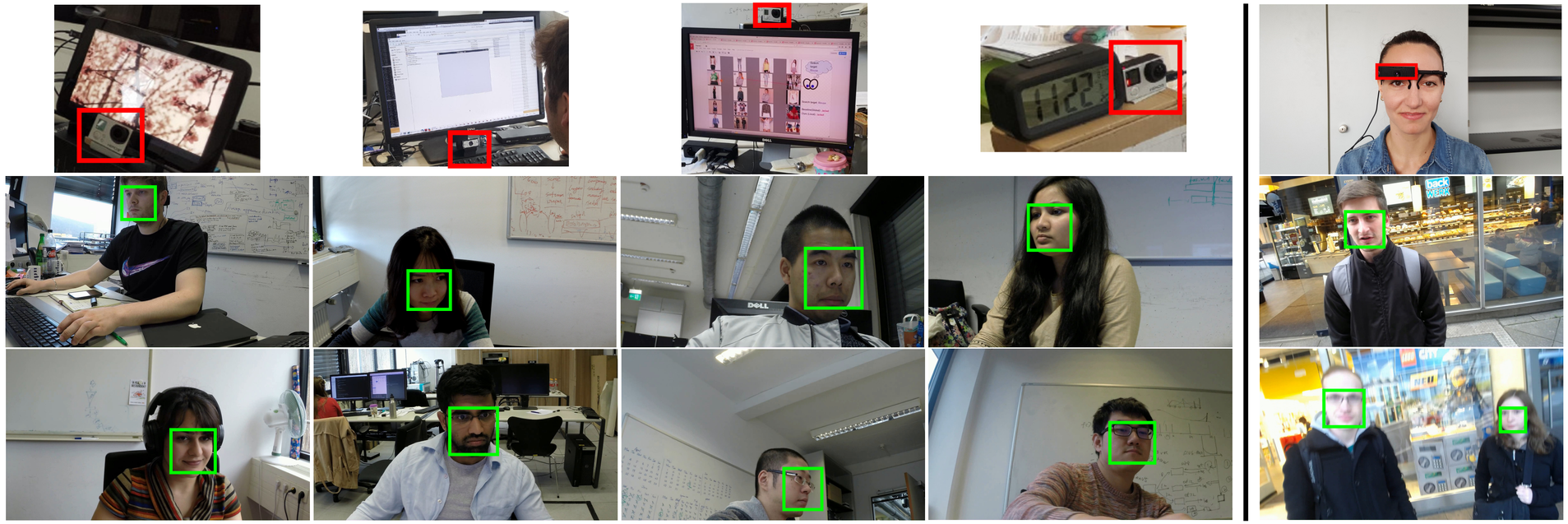
- 対象物体とカメラの位置関係が既知であれば視線推定より容易
- 事前学習することはほぼ不可能

アイコンタクト検出用学習データの自動獲得



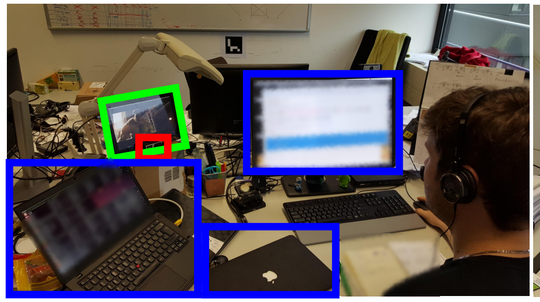
- 注視クラスタの分布から対象物体を発見する
  - アピアランスベース推定の結果は相対的な注視位置の変化はある程度信頼できる
  - 人間は物体領域の中心に注目する傾向があるため、視線推定結果には自然に物体に対応するクラスタができる
  - カメラに最も近い物体がアイコンタクト検出の対象である、という仮定

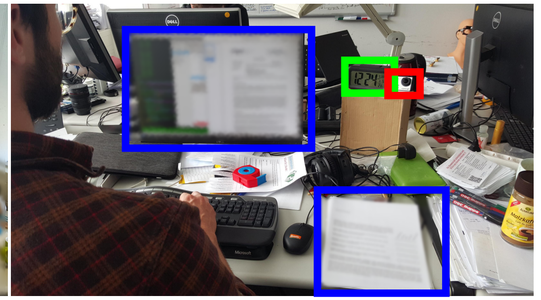
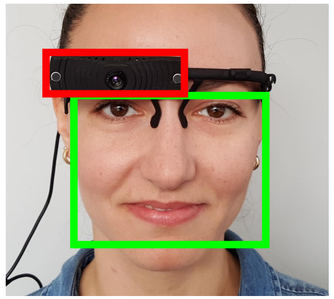
運用・評価実験



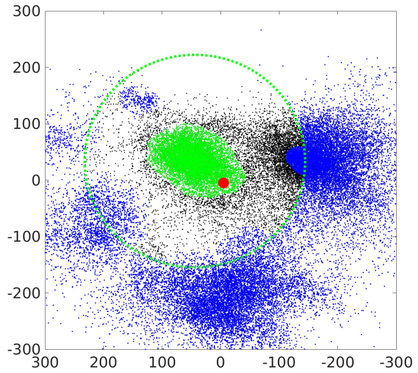
- 二つの運用シナリオでの評価
  - **Object-mounted**: 4通りの物体・カメラ位置の組み合わせ、14人の被験者、各3〜7時間
  - **Head-mounted**: 3人の撮影者がおよそ28人の人物にインタビュー、合計〜5時間
- 真値のアノテーションはテストデータのみ

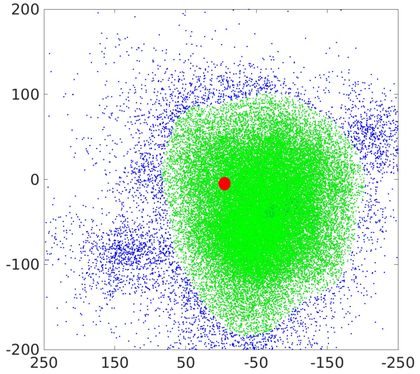
ターゲット検出結果

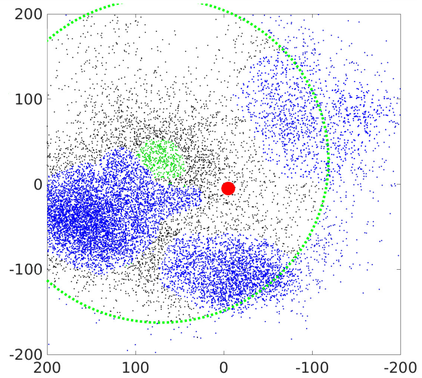


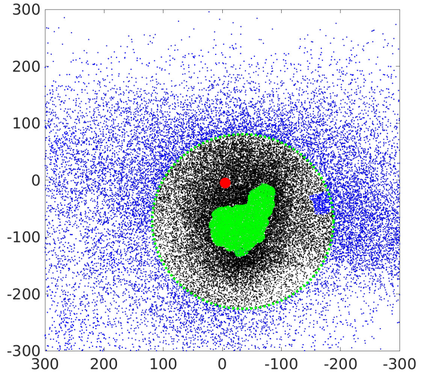



Clustering

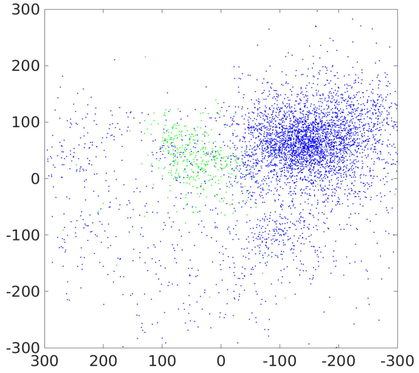


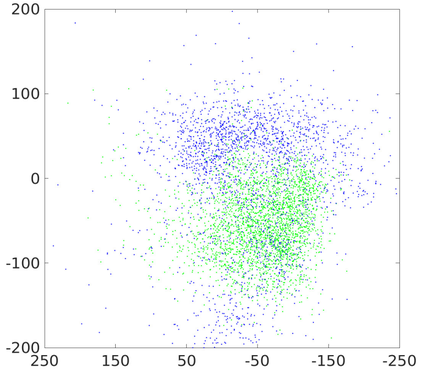


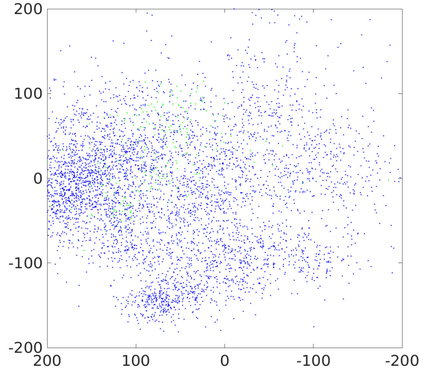


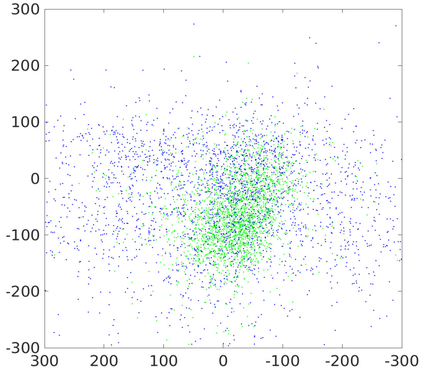


Ground truth

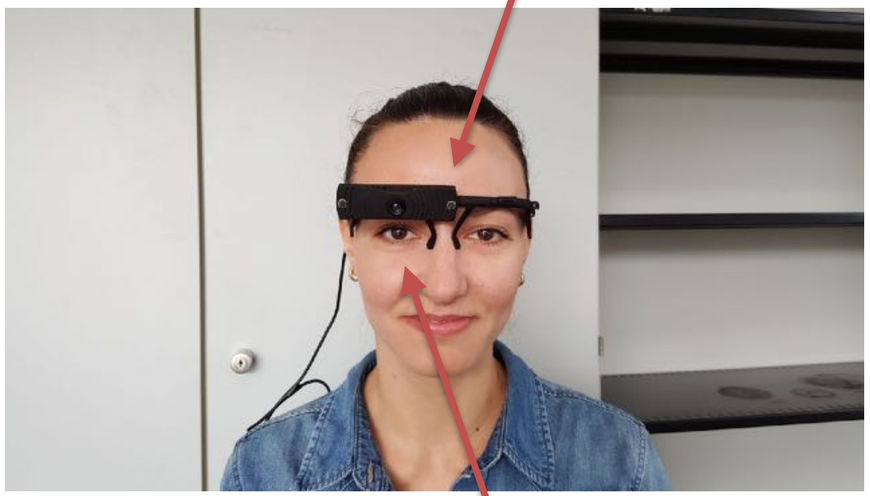


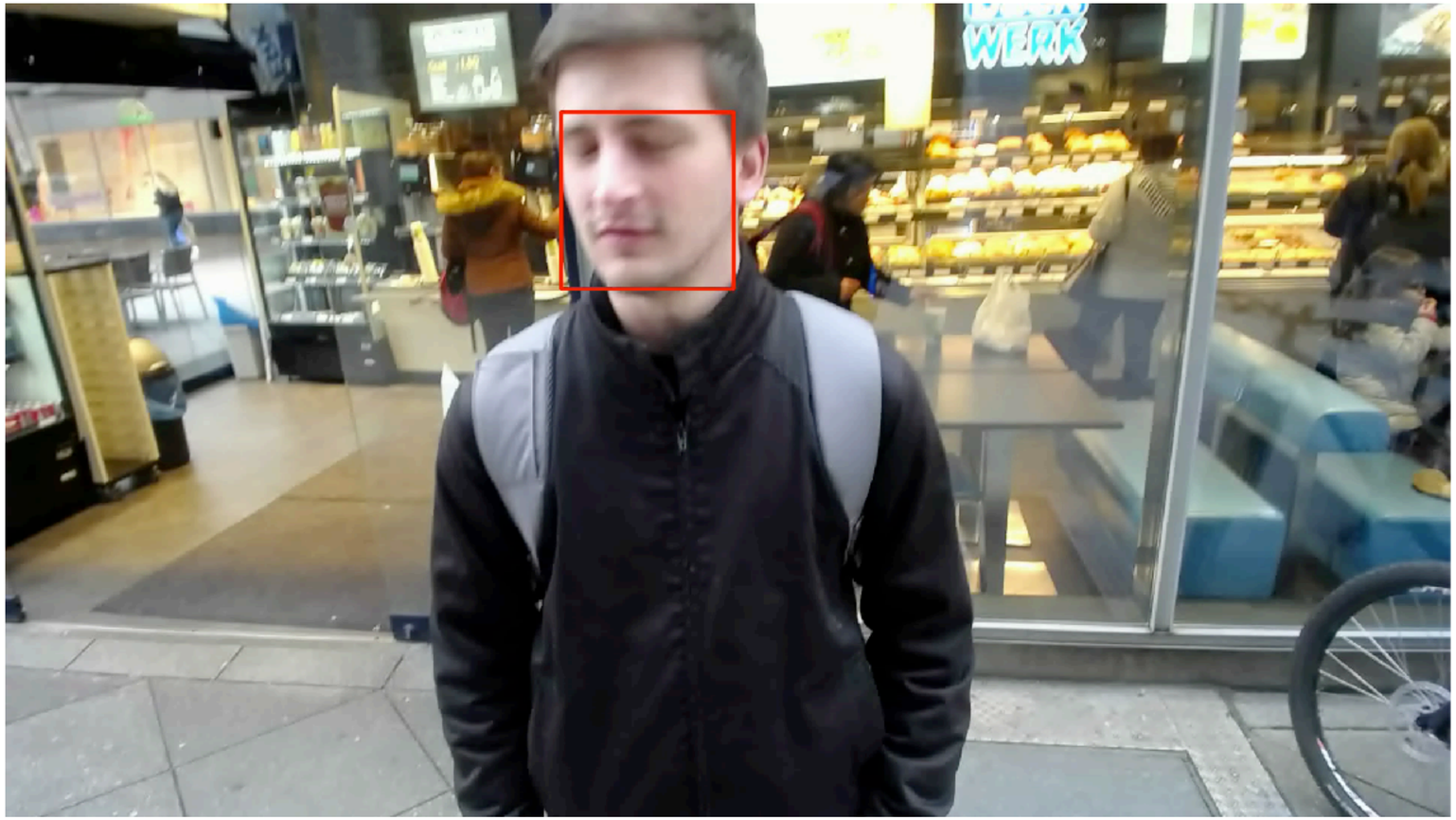




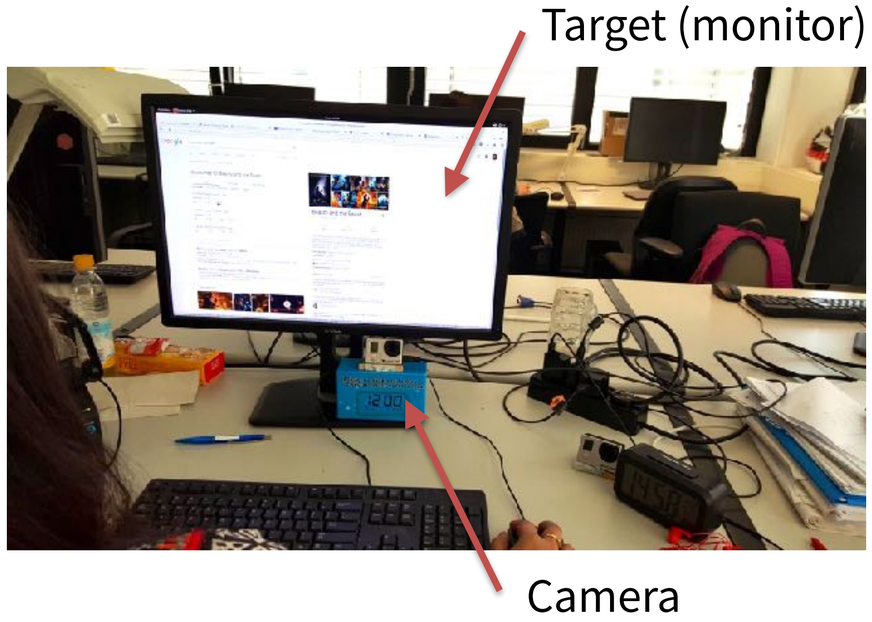


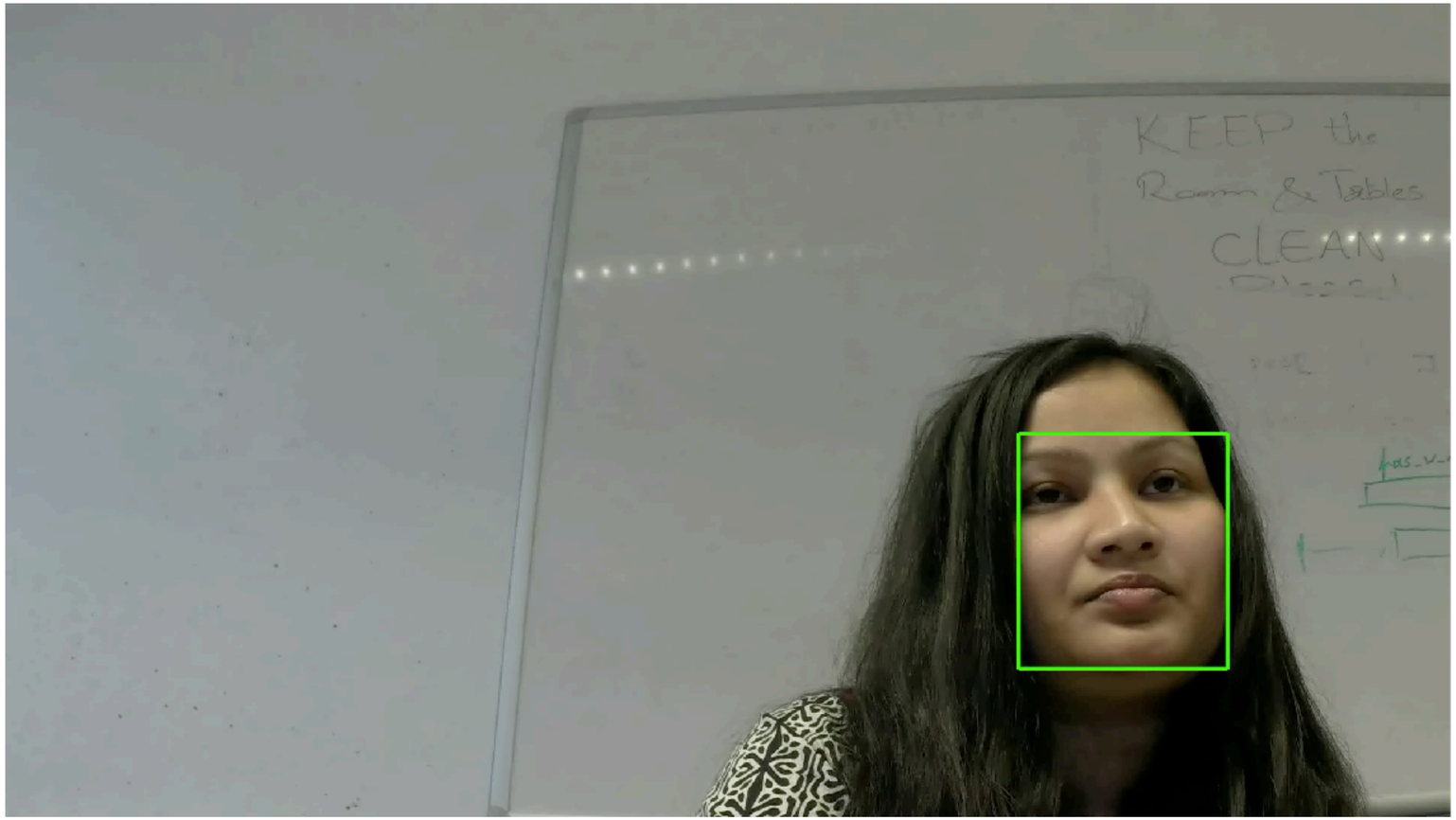
アイコンタクト検出結果





アイコンタクト検出結果





Eye Contact

Non Eye Contact

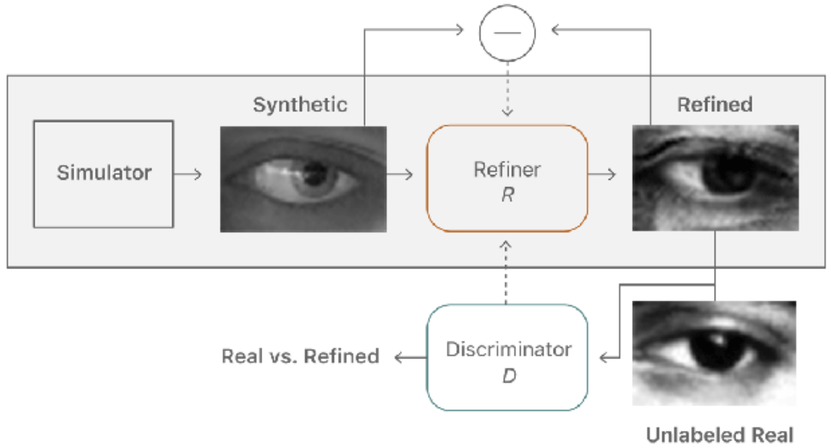
まとめ

- 機械学習アプローチによる視線推定の試み
  - 大規模なデータセットから、人物非依存の視線推定器を学習する
  - 学習・評価データセット構築が重要な研究タスクの一つ
  - 3DCGによる生成画像の利用
  - 顔画像全体を入力特徴とした視線推定のためのCNNアーキテクチャ

- 実環境応用に向けたシステム設計
  - HCIの観点から学習環境とテスト環境の違いに対処する
  - 対象ユーザ・設置環境から自動的に学習データを獲得するための枠組み

今後の展望

- シミュレーション・生成データによる学習
  - 学習データ不足、ドメイン適応の困難さはCV全体でも重要な課題の一つ
  - GANによる画像変換を行う例



[Shrivastava et al., CVPR'17]

- 顔画像からの視線推定
  - 顔全体をCG生成して学習することは可能？
  - 実画像データセットも不足している

- 学習データ獲得のためのシステムデザイン
  - ユーザ・環境の多様性にHCI的な観点から取り組むことは他の認識タスクでも重要
  - 学習アルゴリズムからユーザインタフェースまで一貫した視点で取り組むことで新たな地平が開けるのでは？

ご静聴ありがとうございました

菅野裕介 <sugano@ist.osaka-u.ac.jp>