Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

PROJEKT: GAZE TRACKING
Machine Learning

Letztes Update: 23. Februar 2017

Nadja Kurz

Studienbereich Informatik
Hochschule RheinMain

GLIEDERUNG

Einleitung

ZIELE BZW. INHALTE

- Entwicklung eines CNNs zum Gaze Tracking
- Erkennen von Gesichtsmerkmalen und der 3D-Kopfrotation
- Normalisierung von Bildern einer Webcam
- Entwicklung einer Anwendung zum Sammeln persönlicher Trainingsdaten
- Entwicklung einer lauffähigen Demo
  - Darstellen der Blickrichtung auf dem Bildschirm
  - Sammeln von persönlichen Trainingsdaten
  - 'Zusatztraining' mit den persönlichen Daten

DATENQUELLEN

- Appearance-based Gaze Estimation in the Wild by X. Zhang, Y. Sugano, M. Fritz and A. Bulling
  - CNN-Architektur
  - Normalisierungsansatz
- Learning-by-synthesis for appearance-based 3d gaze estimation by Y. Sugano, Y. Matsushita and Y. Sato
  - Normalisierung
  - Sammeln von eigenen Trainingsdaten (Anwendung)

Datengrundlage

MPIIGAZE

- 213.659 Bilder von 15 verschiedenen Personen
- Bilder wurden in 'natürlicher' Umgebung aufgenommen
- Zeitraum: 3 Monate
- verschiedene Kopfpositionen / Ausrichtungen des Kopfes

Abbildung: Einige Bsp. der normalisierten Bilder des MPIIGaze Datensets.

MPIIGAZE

- (Relevante) Inhalte des Datensets:
  - Original-Bilder (nur Ausschnitte der Augenpartien)
  - Normalisierte Bilder einzelner Augen (60x36 Pixel)
  - Kopfrotation in 3D-Koordinaten
  - Groundtruth in Form von 3D-Koordinaten der 'gaze direction'

Abbildung: Einige Bsp. der normalisierten Bilder des MPIIGaze Datensets.

PERSÖNLICHE TRAININGSDATEN

- Idee: das auf den MPIIGaze-Daten trainierte CNN mit persönlichen Daten 'verfeinern'
- 237 verschiedene Bilder von mir selbst
- unterschiedliche Umgebungen zu unterschiedlichen Zeiten
- Einschränkung: Keine Rotation des Kopfes erlaubt
  - Augen auf einer Höhe mit der Kamera in einem 90° Winkel

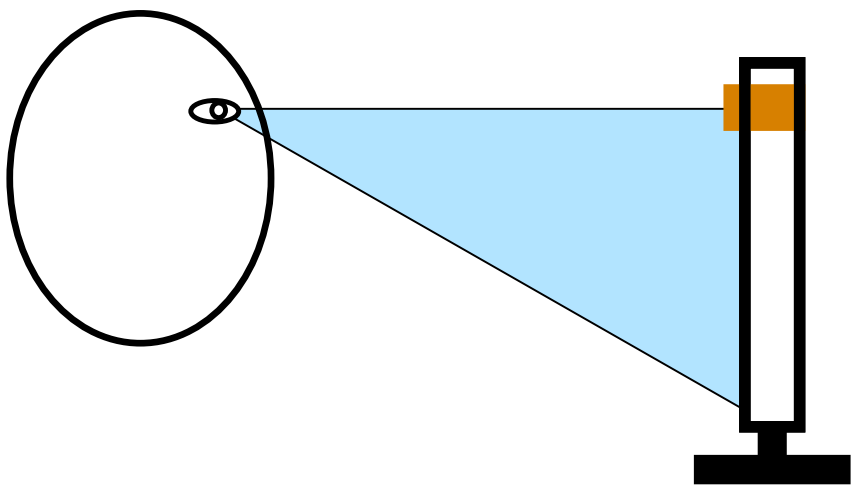


Abbildung: Schematische Darstellung der Kopfposition in Relation zur Kamera.

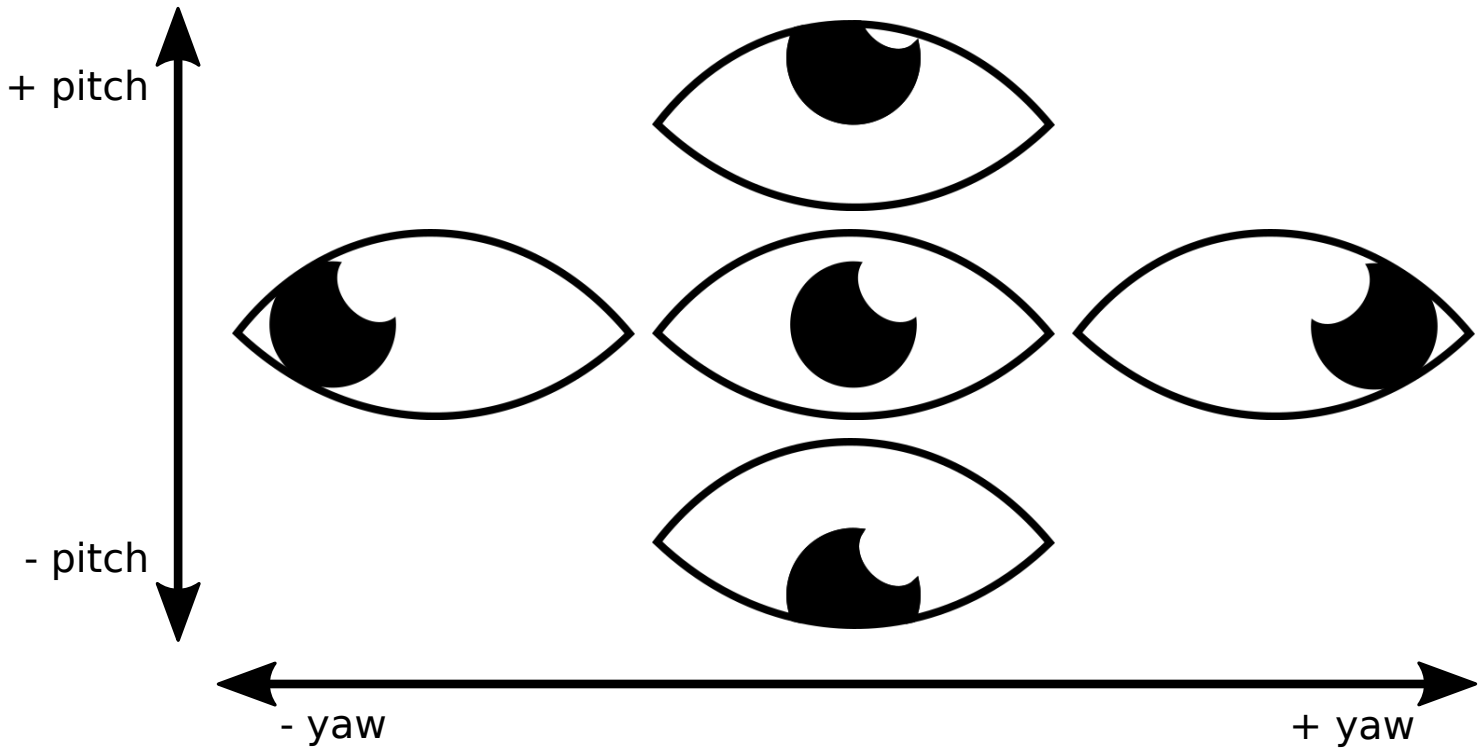
PERSÖNLICHE TRAININGSDATEN

- Weitere Einschränkungen:
  - Halbwegs beleuchtete Umgebung
  - Keine Brillen oder andere Objekte vor den Augen
  - Nur Bilder von einem Auge (hier: das Linke)

Abbildung: Einige Bsp. der normalisierten persönlichen Trainingsbilder.

Architektur und verfolgte Ansätze

PITCH UND YAW



NETZARCHITEKTUR

- Eingabe: Normalisiertes Bild eines linken/rechten Auges
- Ausgabe: 2 Winkel - Pitch und Yaw
  - Pitch: Drehung des Augapfels um die horizontale Achse
  - Yaw: Drehung des Augapfels um die vertikale Achse

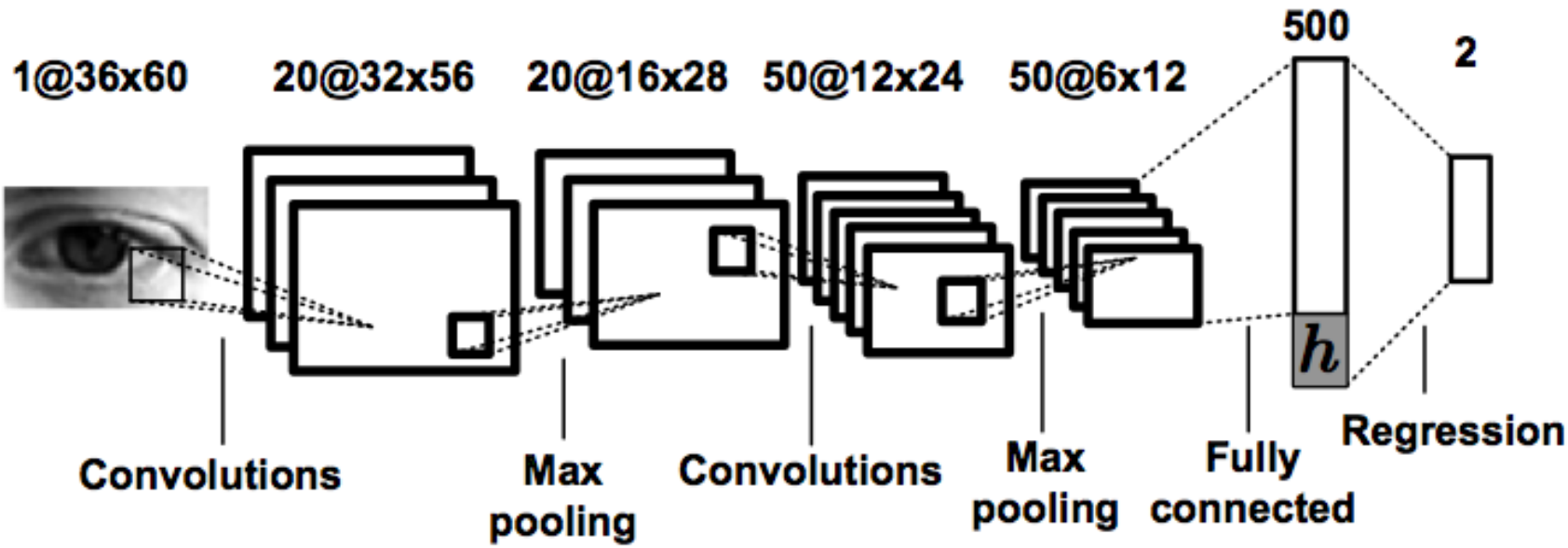


Abbildung: Netzwerkarchitektur des implementierten CNNs [Source: [1]]

[1] Appearance-based Gaze Estimation in the Wild, X. Zhang, Y. Sugano, M. Fritz, A. Bulling

FREIE PARAMETER

- Optimization-Algorithmus: AdamOptimizer
  - weitere freie Parameter geändert: beta1, beta2, epsilon
  - Durch Trial & Error bestimmt
- Loss-Funktion: Summe des quadratischen Fehlers
  - Wichtig! Summe für Pitch und Yaw getrennt $\rightarrow$ 2 Losses
  - Optimierung auf beiden Werten parallel

$$\sum_{i=0}^{n}(output\_pitch - target\_pitch)^2$$

$$\sum_{i=0}^{n}(output\_yaw - target\_yaw)^2$$

FREIE PARAMETER

- Weitere freie Parameter zur Netzarchitektur:
  - Initialisierung der Gewichte:
    - Convolutional-Layers: Zufällige Werte der Normalverteilung
    - FC-Layers: Xavier-Initialisierung[2]
  - Initialisierung der Biases: Konstant 0.1

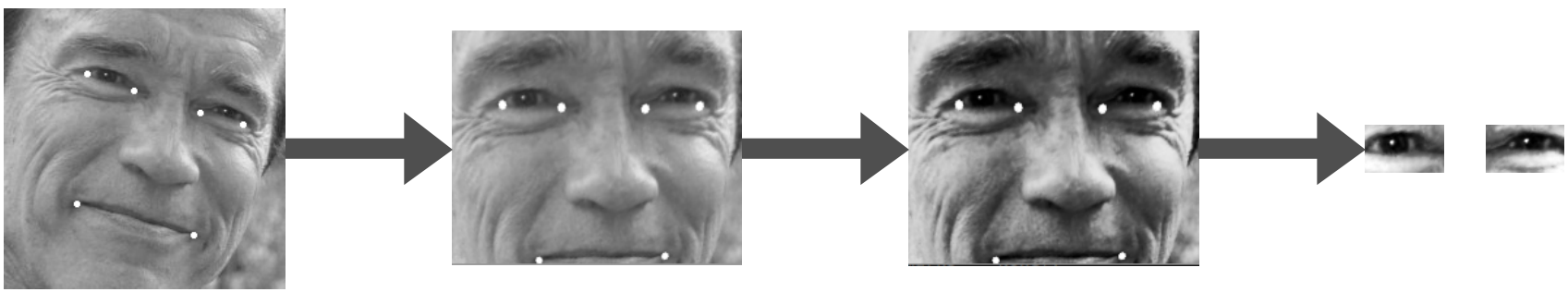
- Allgemeine Parameter zum Training des Netzes:
  - Batchsize:
    - Training auf MPIIGaze: 1000
    - Training auf persönlichen Daten: 100
  - Learningrate: $1e-4 = 0.0001$
  - Anzahl Iterationen:
    - Training auf MPIIGaze: 30.000
    - Training auf persönlichen Daten: $\leqslant 1.000$

[2] Algorithmus, der die Gewichte so initialisiert, dass diese 'gerade so passen'.

Normalisierung

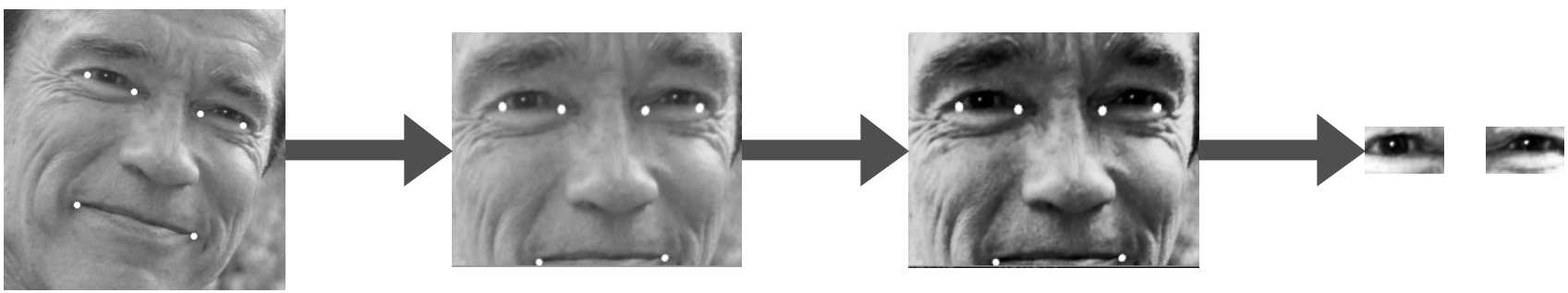
NORMALISIERUNG

- Hintergrund: Augen sehen je nach Kopfposition anders aus
- Idee: Transformation des Gesichtes, sodass Kamera direkt vor den Augen positioniert ist
  - Perspektivische Transformation (Skalierung und Rotation)
- Histogramm-Normalisierung: Reduzierung der Auswirkungen verschiedener Lichtverhältnisse



NORMALISIERUNG - ABLAUF

1. Erkennung von Landmarks
2. Berechnung von Kopfrotation und -transformation (solvePnP)
   - In Relation zu einem generischen 3D-Kopf-Modell
3. Berechnung einer 'Homography' Matrix
4. Perspektivische Transformation
5. Histogramm-Normalisierung
6. Ausschneiden und Skalierung der Augen



Experimente

GÜTEMASS

Durchschnittliche Abweichung [Grad]

$$\frac{1}{\#Samples} \sum |target\_yaw - actual\_yaw|$$

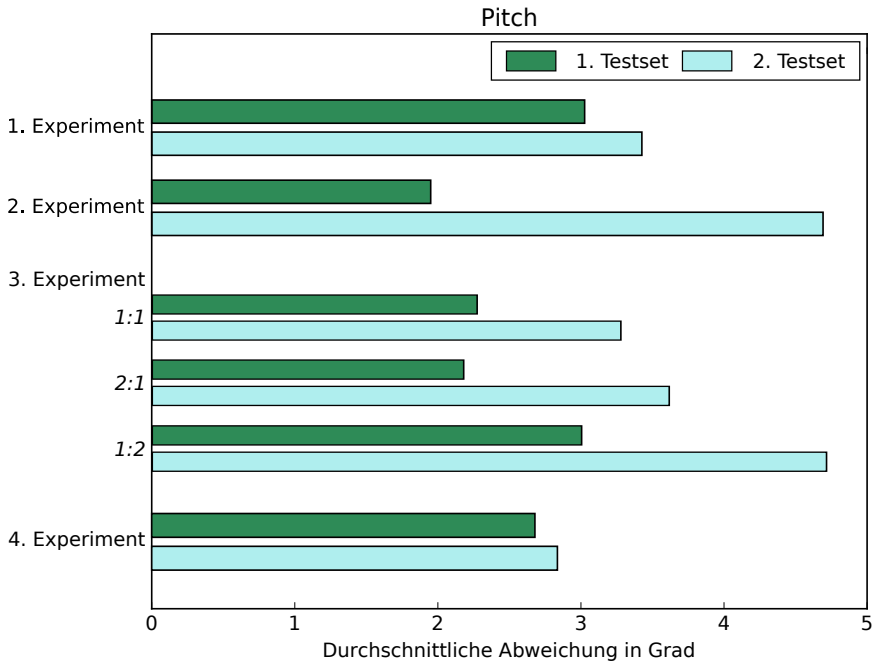
$$\frac{1}{\#Samples} \sum |target\_pitch - actual\_pitch|$$

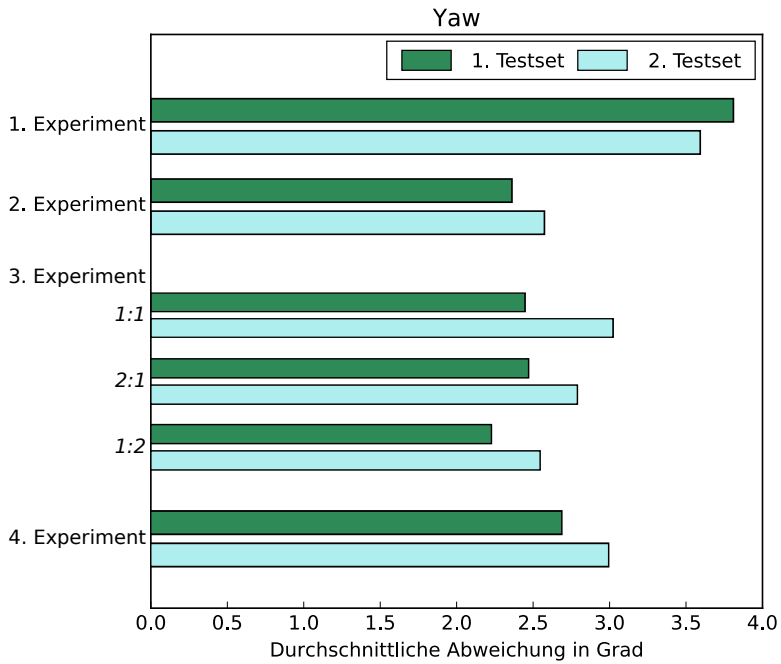
ÜBERSICHT: EXPERIMENTE

Experiment	Vortraining	Zusatztraining
1	MPIIGaze	—
2	MPIIGaze	Nur eigene Daten (227 Bilder)
3	MPIIGaze	Eigene Daten (227 Bilder) + MPIIGaze (x Bilder) Daten
4	Eigene Daten	—

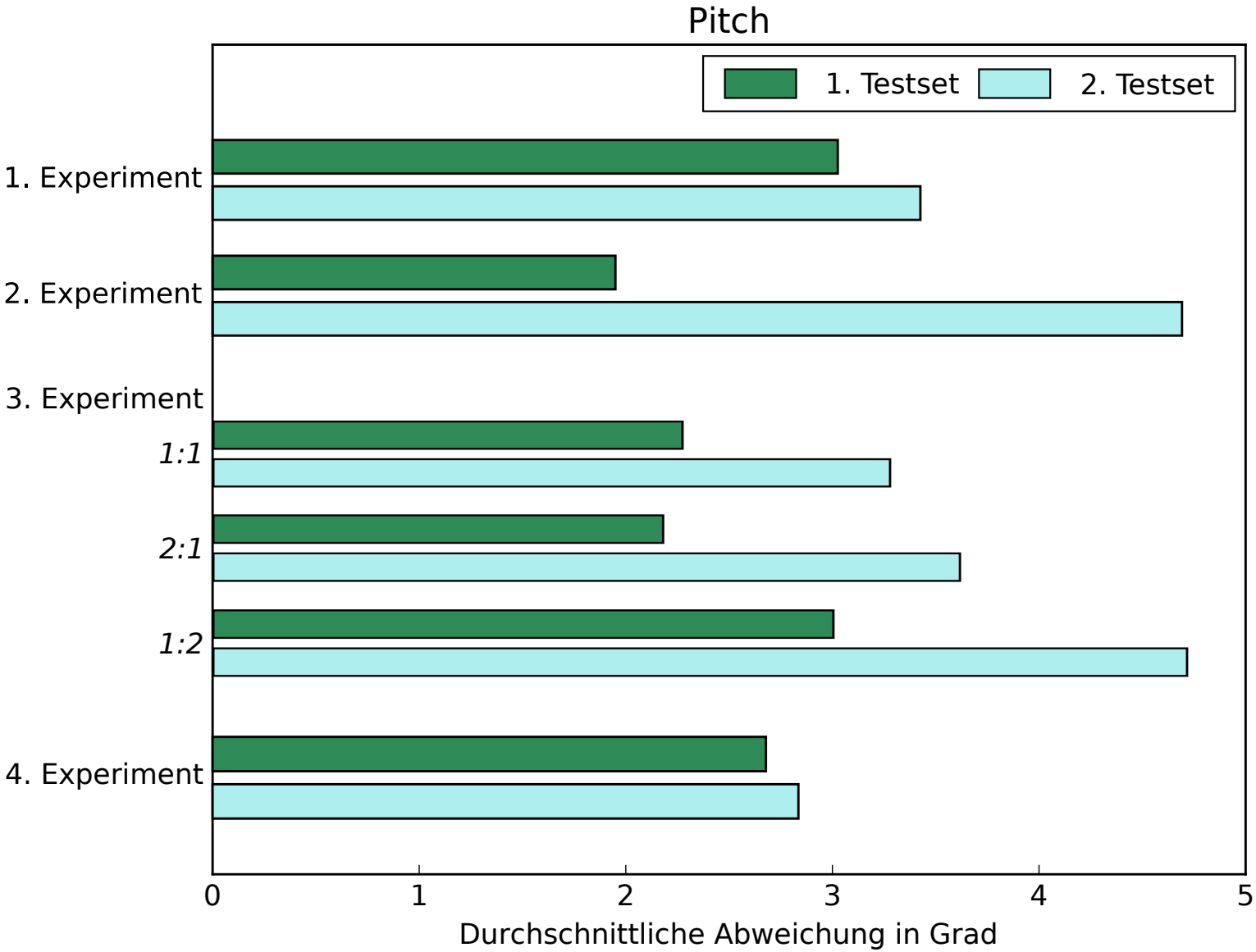
Abbildung: Links: 1. Testset (52 Bilder), Rechts: 2. Testset (90 Bilder)

ERGEBNISSE

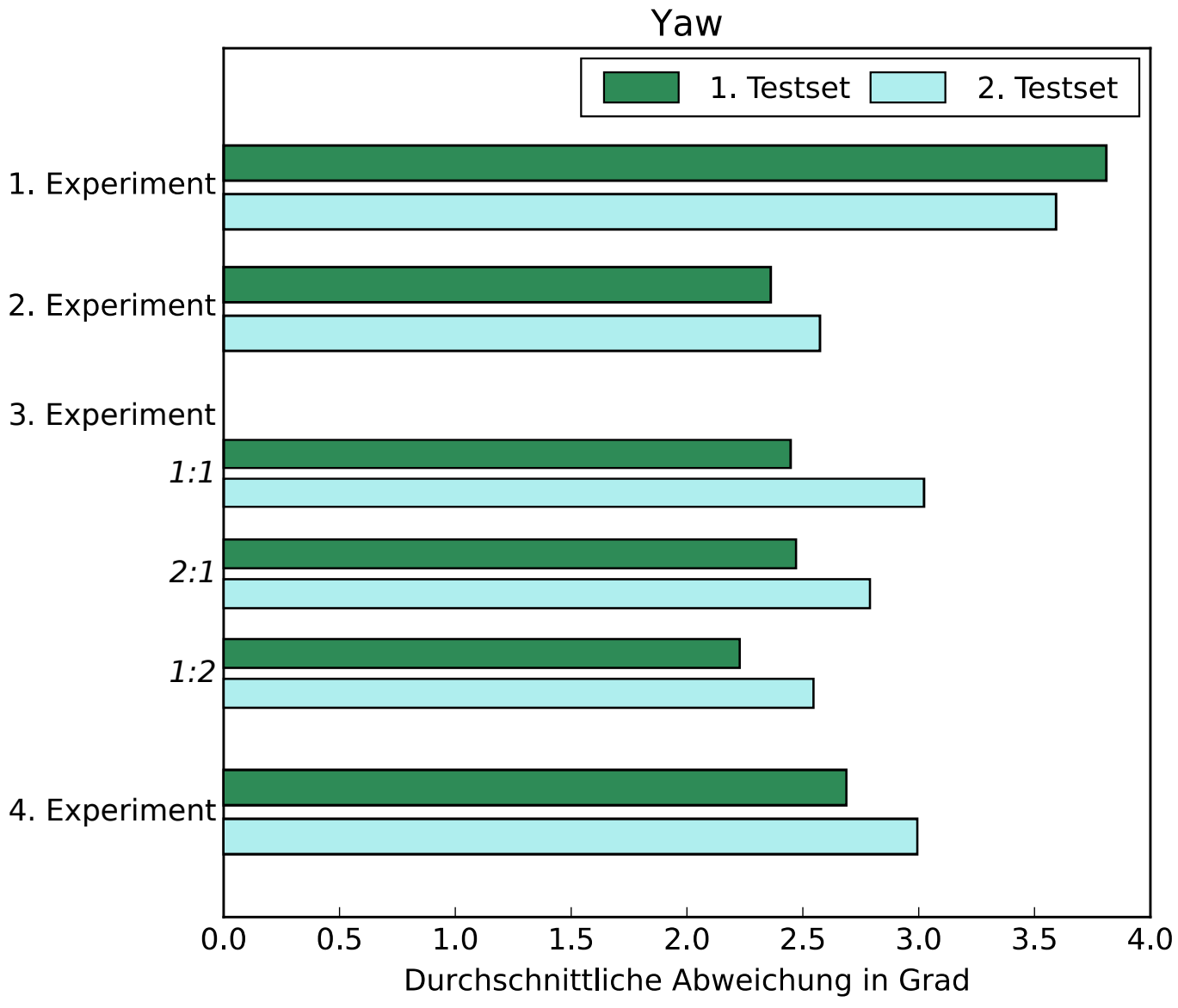

Pitch
1. Testset
2. Testset
1. Experiment
2. Experiment
3. Experiment
1:1
2:1
1:2
4. Experiment
0
1
2
3
4
5
Durchschnittliche Abweichung in Grad


Yaw
1. Testset
2. Testset
1. Experiment
2. Experiment
3. Experiment
1:1
2:1
1:2
4. Experiment
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Durchschnittliche Abweichung in Grad

ERGEBNISSE - PITCH



ERGEBNISSE - YAW



ERGEBNISSE - ZUSAMMENFASSUNG

- MPIIGaze vs persönliche Daten:
  - Persönliche Daten verbessern Qualität
  - Eventuell abhängig von der Normalisierung
  - MPIIGaze + Zusatztraining = gute Alternative
- Zusatztraining:
  - Kann Generalisierungsfähigkeit beeinträchtigen
  - Verbessert Qualität für einzelne Person deutlich
- Pitch ist schwieriger zu Klassifizieren als Yaw
- Insgesamt: Durchschnittliche Abweichung $< 5°$

Fazit

FAZIT UND AUSBLICK

- MPIIGaze Daten sind teilweise nicht brauchbar
- Normalisierung wurde in dem Originalpaper kaum erläutert
  - Viel 'herumprobieren' und nachforschen in Quellen des Papers
- Training hat nicht mit den gleichen Parametern des CNNs des Originalpapers funktioniert (andere Learningrate, etc.)

Abbildung: Ausschnitt einiger problematischer Bilder des MPIIGaze Datensets.

FAZIT UND AUSBLICK

- Brauchbare Lossfunktion zu bestimmen war größte Hürde
  - Optimierung mit kombiniertem Losswert hat nicht funktioniert
  - Lossfunktion, die im Originalpaper verwendet wurde, hat Fehlermeldungen erzeugt
- Weitere Ansätze, die nicht umgesetzt wurden:
  - Andere CNN Architektur bzw. komplexere
    - Mehr Convolutional Layer
    - Nicht nur max Pooling

Vielen Dank für Ihre Aufmerksamkeit

QUELLEN

- Appearance-based Gaze Estimation in the Wild, X. Zhang, Y. Sugano, M. Fritz, A. Bulling
- Learning-by-synthesis for appearance-based 3d gaze estimation by Y. Sugano, Y. Matsushita and Y. Sato
- Head Pose Estimation. Available from http://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/
- Dlib - Landmark detection. Available from https://matthewearl.github.io/2015/07/28/switching-eds-with-python/
- Perspective Transformation. Available from http://stackoverflow.com/questions/23275877/opencv-get-perspective-matrix-from-translation-r
- Special Olympics Österreich. Arnold Schwarzenegger. Available from https://www.flickr.com/photos/so-austria/19948240828/ License.