Stabile Gesichtserkennung mittels Deep Learning

Jan Kaßel

Matrikelnr. 3724135
M.Sc. Informatik
1. Fachsemester

31. März 2018

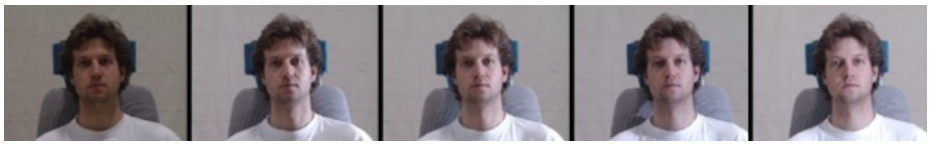
Forschungsseminar Deep Learning
Wintersemester 2017/2018
Universität Leipzig

Zusammenfassung. Die Erkennung und Beschreibung von Gesichtern ist ein gefragtes Feld der Computer Vision, das unter anderem Überwachungs- und Sicherheitssysteme zu seinen prominenten Nutzern zählt. Die Varianz der Daten, insbesondere bei Betrachtungen von Gesichtern unter verschiedenen Blickwinkeln und Beleuchtungen, birgt allerdings Schwierigkeiten für spezialisierte, weniger flexible Ansätze zur Gesichtserkennung. Zhu et al. zielen in ihrer Forschung von 2013, "Deep Learning Identity-Preserving Face Space", darauf ab, mit Hilfe eines künstlichen neuronalen Netzwerks bestimmte posen- und beleuchtungsunabhänge Merkmale in Gesichtern zu extrahieren. Mit Hilfe dieser Merkmale, *Face Identity-Preserving Features* (FIP) genannt, lassen sich Gesichter in einer optimierten Frontalansicht rekonstruieren. Mit Hilfe dieser Ansicht werden Beleuchtung und Blickwinkel nahezu neutralisiert, womit Erfolgsraten bei der Gesichtserkennung bestehender Systeme deutlich verbessert werden können. Zhu et al. zeigen während eines Experiments im Vergleich mit gängigen Ansätzen, dass die von ihnen konzipierten FIP und rekonstruierten Ansichten Gesichter zuverlässige Repräsentationen darstellen, die eine Erkennung von Gesichtern unter verschiedenen Blickwinkeln mit hohem Erfolg ermöglicht.

1 Einführung

Während sich grundlegende Forschung der Computer Vision in den frühen 1990er Jahren noch mit der prinzipiellen Extraktion von Gesichtsmerkmalen innerhalb eines Bildes beschäftigte [2], sind zeitgenössische Systeme bereits in der Lage, menschliche Gesichter problemlos zu erkennen, ihre Mimik zu deuten und sie eindeutig Personen zuzuordnen [1]. Dies ist nicht zuletzt aufgrund effizienterer Technologie und gewachsener Rechenleistung möglich, die die damalige bei weitem übertrifft. Anwendungen der Gesichtserkennung, beispielsweise Videoüberwachung oder Systeme zur Identifizierung von Personen, profitieren von

(a) Rotation eines Gesichts [4]



(b) Verschiedene Beleuchtung eines Gesichts [4]

Abbildung 1: Einträge der MultiPIE Face Database von Gross et al. wurden unter verschiedenen Beleuchtung und Blickwinkeln aufgenommen, um einen Vergleichsbasis für Systeme zur Gesichtserkennung zu schaffen [4, 5].

diesem Fortschritt: Kürzlich führte die Veröffentlichung des Systems *Face ID* der Firma Apple vor, dass solche Technologien längst zuverlässig arbeiten und marktreif sind [3].

Grundsätzlich lässt sich die Arbeitsweise von Systemen zur Gesichtserkennung in zwei Ansätze unterteilen. Wiskott et al. konzipierten schon früh eine Methode, um individuelle Merkmale im Gesicht erkennen und mittels Deskriptoren bestimmten Identitäten zuzuweisen [6]. Es wird versucht, mittels Analyse der Textur von Gesichtsaufnahmen bestimmte Merkmale im Gesicht zu extrahieren und deren Ausprägungen festzuhalten. Diese Merkmale erweisen sich je nach Beschaffenheit als instabil gegenüber leichten Drehungen des Kopfes oder Veränderungen in der Beleuchtung (Abbildung 1). Die Arbeit mit dreidimensionalen Modellen von Gesichtern erleichtert dies mittels Transformationen im dreidimensionalen Raum, birgt allerdings eine aufwendige Umsetzung und Fehleranfälligkeit durch 3D-Scans und die zusätzliche räumliche Dimension [1]. Li et al. hingegen beziehen zum Vergleich zwar eine Datenbank von dreidimensionalen Gesichts-Scans, ermöglichen allerdings die Erkennung von Gesichtern in zweidimensionalen Daten: Anhand von Fragmenten wird aus einem Bild eine dreidimensionale Repräsentation des Gesichts berechnet, die dann mit Einträgen aus der Datenbank verglichen wird [1, 7]. Im dreidimensionalen Raum reduzieren sich Probleme der Kopfrotation lediglich auf Anwendungen linearer Algebra, die sich aufgrund ihrer linearen Beschaffenheit trivial lösen lassen [1].

In dieser Arbeit setzen wir uns mit einem System auseinander, das Zhu et al. in 2013 konzipiert haben. Ihr Ansatz sieht vor, die Problematik Gesichtserkennung, insbesondere im Hinblick auf Invarianz gegenüber verschiedener Posen

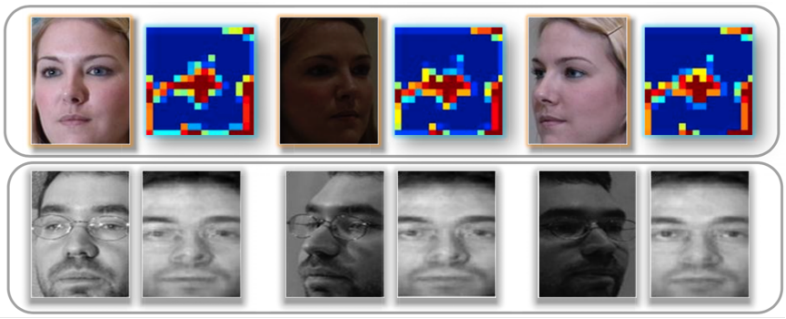


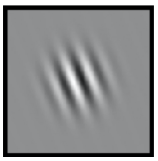
Abbildung 2: Repräsentation der Gesichter zweier Personen (jeweils obere und untere Reihe) als zweidimensionaler Deskriptor (*Face Identity-Preserving Features*, oben) und rekonstruiertes in neutraler Pose Gesicht (*Canonical View*, unten) [1].

und Beleuchtungen, mit maschinellem Lernen anzugehen. Ein speziell konzipiertes *Deep Neural Network* bzw. *Convolutional Neural Network* (CNN) wurde so entworfen, dass über mehrere *Hidden Layers* hinweg Merkmale des menschlichen Gesichts erkannt werden, unabhängig von der Beleuchtung oder leichten Drehungen des Kopfes. Das Netzwerk erwartet als Eingabe zweidimensionale Bilder. Die Invarianz der Merkmale gegenüber Beleuchtung und Blickwinkel wurde bei der Architektur des Netzwerks von Zhu et al. bewusst konstruiert [1], wie in Kapitel 3 erläutert wird.

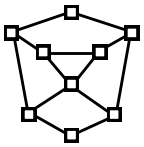
Diese Merkmale, *Face Identity-Preserving Features* (FIP) genannt, sind die Ausgabe einer Aneinanderreihung von Schichten in dem neuronalen Netzwerk und lassen sich mit Hilfe des Netzwerks in eine neutrale Darstellung des Gesichts überführen, der *Canonical View*, wie in Abbildung 2 illustriert. Diese Ansicht stellt das Gesicht frontal und in neutraler Beleuchtung da, was Varianzen unter Rekonstruktionen von Aufnahmen aus verschiedenen Blickwinkeln auf ein Minimum reduziert, wenn nicht sogar vollkommen neutralisiert [1].

Die Architektur des Netzwerks erstreckt sich über vier Blöcke, die von Elemente der CNNs verwenden: Die Blöcke setzen sich aus aus *Locally Connected*, *Pooling* und *Fully Connected* Layers zusammen. Durch diesen Aufbau können Features verschiedener Art erkannt, betont oder entfernt werden und haben aufgrund der Tiefe der Schichten und der Eigenheiten von Convolutional Layers hohe semantische Aussagekraft, sind zusätzlich sogar verhältnismäßig gut lernbar: Convolutional Layers verwenden viele Filtern mit kleinen rezeptiven Feldern, die durch dünn besetze Matrizen mit wenigen Gewichten repräsentiert werden. Ngiam et al. beschreiben dies ausführlich [8], wie in Kapitel 2 dargelegt wird.

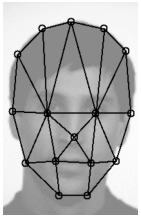
Im Folgenden setzen wir uns zunächst mit Forschung auseinander, die bereits im Feld der Gesichtserkennung mit neuronalen Netzwerken getätigt wurde. In



(a) *Gabor Wavelet* [6]



(b) Abstrakter *Bunch Graph* [6]



(c) *Bunch Graph*, berechnet für eine Gesichtsaufnahme [6]

Abbildung 3: Funktionsweise des Gabor-Deskriptors. Mittels *Convolution Filters* werden *Gabor Wavelets* (Abbildung 3a) gesammelt, bezeichnet *Jet-Samples*, die als Einheit ein Merkmal im Gesicht beschreiben. Diese Jets werden mittels *Bunch Graphs* Positionen im Gesicht zugeschrieben (Abbildung 3b) [6].

Kapitel 2 werden artverwandte Ansätze erklärt, die die Forschung von Zhu et al. beeinflusst haben oder Grundlegendes geschaffen haben. Des Weiteren gehen wir auf Methoden ein, die zur Erkennung der Face Identity-Preserving Features beitragen. Kapitel 3 wird sich näher mit besagter Architektur des Netzwerks von Zhu et al. befassen. Insbesondere die mathematische Repräsentation des Netzwerks unter Berücksichtigung der Eigenheiten des Convolutional Neural Network sind Diskussionspunkt. Das Training des Netzwerks wird in Kapitel 4 beschrieben, und in Kapitel 5 und 6 werden Experimente von Zhu et al. und deren Auswertung erläutert.

2 Forschung

Wiskott et al. bauen ihren Ansatz zur Gesichtserkennung auf Graphen auf, sogenannten *Bunch Graphs*, um Merkmale des menschlichen Gesichts in Relation zu ihrer Position im Gesicht zu bringen (Abbildung 3b, 3c) [6]. Diese Struktur birgt entscheidende Vorteile zur Erkennung von Gesichtern unter verschiedenen Blickwinkeln: Ein Bunch Graph lässt sich auf triviale Weise linear transformieren und macht lediglich Annahmen über die Struktur des Gesichts, nicht aber über die Beschaffenheit dessen Merkmale. Um die jeweiligen Merkmale zu erkennen, werden sie in der Bildtextur mithilfe eines Gabor-Filters aufgenommen und in *Jet Samples* kombiniert (Abbildung 3a). Graph und Jet Samples werden auf einem Datenbestand trainiert, der Bilder von Gesichtern aus verschiedenen Blickwinkeln beinhaltet. Somit werden zum einen abstrakte Bunch Graphs erstellt, zum anderen die Jet Samples zu Merkmalen der jeweiligen Person. Die Zuweisung der Jet Samples zu den Graph-Knoten ist flexibel, jedoch lässt dieser Ansatz aufgrund der limitierten Aussagekraft der Jet Samples nur die Erkennung von

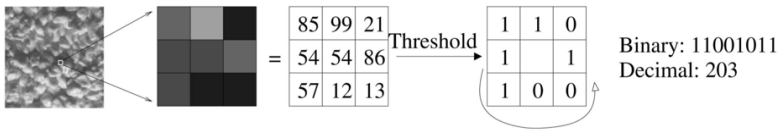


Abbildung 4: Anwendung des *Local Binary Pattern* Deskriptors [9]. Eingabedaten werden in kleinen Ausschnitten analysiert und mit Hilfe von Thresholding encodiert. Die resultierenden Werte beschreiben in Kombination mit anderen charakteristische Merkmale und können mittels eines Histogramms dargestellt werden.

Gesichtern zu, die zuvor trainiert wurden [6].

Ähnlich setzen Ahonen et al. in dem von ihnen konzipierten Gesichts-Deskriptor auf eine Filter-basierte Texturerkennung [9]. Texturelle Merkmale, *Local Binary Patterns* (LBP), werden über Filter mit kleinen rezeptiven Feldern aus den Eingabedaten aufgenommen, via *Thresholding* encodiert und schließlich in einem dreidimensionalem Histogramm visualisiert (Abbildung 4). Diese Art der Codierung macht die gesammelten Features invariant gegenüber Rotationen und sie somit geeignet für Erkennung von Gesichtern aus verschiedenen Blickwinkeln [10, 9]. Ahonen et al. kombinieren die lokalen Features in bestimmten Regionen innerhalb des Gesichts, um globale, semantische Deskriptoren zu erhalten.

Im Gegensatz zu den Textur-basierten Deskriptoren von Wiskott et al. und von Ahonen et al., haben Cao et al. einen Deskriptor entwickelt, *LE* genannt, dessen Merkmale mittels *Deep Learning* gewonnen werden. Dynamische und vielseitige Daten benötigen auch komplexe Deskriptoren, die mittels manuell konzipierter Modelle nicht abgedeckt werden können [11]. Mittels Machine Learning wurde ein Modell erzeugt, das dynamisch, auf Trainingsdaten basierte Merkmale 'lernen' kann. Diese Merkmale können schließlich durch *Support Vector Machines* zur Erkennung von Gesichtern eingesetzt werden kann. Eine von Cao et al. anschließend durchgeführte Studie zeigt, dass die erlernten Deskriptoren im Vergleich zu Gabor- und LBP-Deskriptoren auf Basis des MultiPIE-Datensatzes [5] bis zu 10% bessere Ergebnisse erzielen können.

Convolutional Neural Networks (CNNs) ermöglichen unter anderem, durch Einstellung der Gewichte jeweiliger Convolution-Filter Varianzen zu neutralisieren: Manuelles oder automatisiertes Training kann die entsprechenden Gewichte der künstlichen Neuronen so adjustieren, dass bestimmte Datenregionen bevorzugt, andere benachteiligt werden. Da Features unabhängig von ihrer Position in den Daten detektiert und mittels Pooling verstärkt werden, sind sie natürlicherweise invariant gegenüber Verschiebungen [8]. Ngiam et al. stellen den Ansatz vor, CNNs mittels *Tiling* auch gegenüber komplexerer Invarianzen zu rüsten, wie Rotation und Skalierung [8]. Das Tiling sieht vor, Gewichte zwi-

schen Convolution-Units innerhalb eines Layers nur dann in Relation zu bringen (*tying*), wenn sie eine bestimmte Distanz zueinander einhalten. Durch den flexiblen Einsatz von Pooling-Units gemeinsam mit Convolution-Units können deutlich komplexere Features erlernt und erkannt werden. Die nur wenigen Gewichte erhöhen deren Anzahl lediglich um einen linearen Faktor und ermöglichen effizientes Training des Netzwerks [8].

3 Netzwerk-Architektur

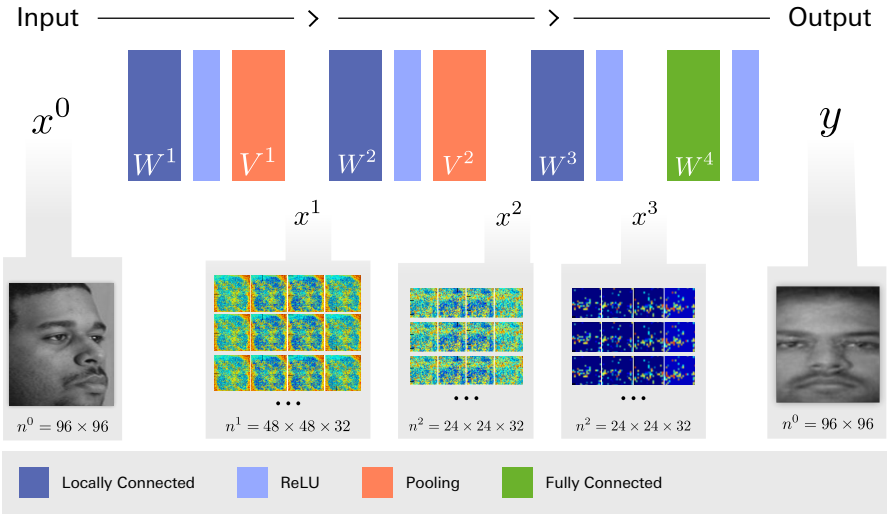


Abbildung 5: Architektur des künstlichen neuronalen Netzwerks [1]. Es erhält als Eingabe x^0 ein Bild, gibt die *Face Identity-Preserving Features* x^3 aus und kann das Gesicht in der *Canonical View* y, einer neutralen Ansicht, rekonstruieren. x^1 und x^2 sind Outputs der ersten beiden lokal verbundenen Schichten, die 32 Feature Maps mit verschiedenen Aktivierungen ausgeben.

Die Übersicht der bisherigen Systeme zeigt, dass Learning-basierte Deskriptoren zur Gesichtserkennung den klassischen, manuell konzipierten Deskriptoren aufgrund ihres vielseitigen semantischen Ausdrucks überlegen sind und abstraktere Merkmale beschreiben können. Allerdings greifen auch traditionelle Deskriptoren auf kleine, flexible *Receptive Fields* für ihre Filter zurück, die, je nach Codierung, besondere invariante Eigenschaften haben können, zum Beispiel gegenüber Rotation und Beleuchtung. Ngiam et al. haben gezeigt, dass *Convolutional Neural Networks* mittels *Tiling* und kleiner Receptive Fields ebenso gut Merkmale erkennen können, die Invarianzen besitzen [8]. Zhu et al. haben sich

dessen in ihrer Konzeption eines neuronalen Netzwerks bedient, wie wir im Folgenden lernen werden.

Das künstliche neuronale Netzwerk setzt sich aus vier Blöcken zusammen. Um lokale und globale Informationen zu extrahieren, werden sowohl *Locally Connected (Convolutional)* als auch *Fully Connected* Layers verwendet [1]. Als Eingabe erhält das Netzwerk das Bild eines Gesichts, x^0, der Größe $n_0 = 96 \times 96 = 9216$. Nach Anwendung der ersten drei Schichten erhält man die *Face Identity-Preserving Features* (FIP), x^3; eine Sammlung an 32 Feature Maps der Größe 24×24, die die Merkmale eines Gesichts beschreiben, unabhängig von Pose und Beleuchtung. Die vierte Schicht rekonstruiert das Gesicht anhand der Features aus x^3 in die *Canonical View*, y, mit $n_0 = 96 \times 96$ Dimensionen. Im Folgenden werden Aufbau und Funktionsweise der jeweiligen Schichten erläutert.

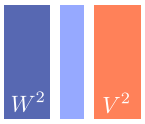
Erster Block

Der erste Block des Netzwerk besteht aus einem *Locally Connected Convolutional Layer* und einem *Max-Pooling Layer*. Ersteres transformiert die Eingabe, x^0, in 32 Feature Maps. Mathematisch betrachtet besteht die Gewichtsmatrix des Layers, W_1, aus 32 dünn besetzten Unter-Matrizen, $W^1 = [W_1^1; W_2^1; ...; W_{32}^1]$, $\forall W_i^1 \in \mathbb{R}^{n_0 \times n_0}$ [1]. Jedes W_i^1 repräsentiert einen Filter, der mit einem Receptive Field von 5×5 auf x^0 angewandt wird. Das Pooling Layer wird als Binärmatrix, V^1, mit $V_{i,j}^1 \in \{0, 1\}$ dargestellt. Es reduziert die Größe der Feature Maps von $n_0 = 96 \times 96$ auf 48×48 Dimensionen, also $n_1 = 48 \times 48 \times 32$. Die Unterteilung des Convolutional Layers auf viele kleine, unabhängige Filter als auch das Downsampling des Pooling Layers folgen dem Ansatz der *Tiled CNNs* nach Ngiam et al. [8]. Insbesondere wird somit die Stabilität gegenüber Rotationen und Bewegungen des Gesichts in der Eingabe verstärkt.
Die resultierenden Feature Maps x^1 besitzen eine Größe von jeweils 48×48 und lassen sich durch Multiplikation mit den Gewichtsmatrizen W^1 und V^1 berechnen [1],

$$x_i^1 = V^1 \sigma(W_i^1 x^0). \tag{1}$$

σ bezeichnet hierbei die *Rectified Linear Unit*, $\sigma(x) = max(0, x)$. Die Verwendung von σ erhält relative Verhältnisse zwischen Features und ist somit stabil gegenüber verschieden starken Ausprägungen von Features [1, 12].

Zweiter Block



Ähnlich zum ersten stellt sich auch der zweite Block dar. Ein Locally Connected Layer, W^2, und ein weiteres Pooling Layer, V^2, werden auf die 32 Feature Maps aus x^1 angewandt. Die Gewichtsmatrix des Locally Connected Layer, W^2 ist ebenso dünn besetzt und besteht aus 32 Unter-Matrizen, $W_i^{2\,32}_{i=1}, \forall W_i^2 = \{W_{i,j}^2\}_{i=0}^{32} \in \mathbb{R}^{48\times48,48\times48}$ [1]. Alle W_i^2 werden auf jede Feature Map aus x^1 angewandt und die Ergebnisse pro Feature Map summiert, damit x^2 erneut aus 32 Feature Maps besteht. W^2 erhält man folglich via $W^2 = [W_1^{2'}; ...; W_3^{2'}2]$, mit $W_i^{2'} = [W_i^2]_{j=1}^{32}$, also der 32-fachen Wiederholung von W_i^2. W^2 projiziert somit jedes x_i^1 wiederum auf jeweils 32 Feature Maps,

$$x_i^2 = \sum_{j=1}^{32} V^2 \sigma(W_{i,j}^2 x_i^1). \tag{2}$$

Um x^2 zu erhalten, wird in dem Pooling Layer, V^2, ein Downsampling für jedes x_i^2 auf 24×24 Dimensionen angewandt, also $n_2 = 24 \times 24 \times 32, x^2 \in \mathbb{R}^{n_2}$. Das resultierende x^2 besitzt somit $n_2 = 24 \times 24 \times 32$ Dimensionen [1].

FIP

Im dritten Block, der lediglich aus einem Convolutional Layer besteht, werden die Face Identity-Preserving Features berechnet. Der Aufbau dieses Layers ähnelt dem von W^2: Die Gewichtsmatrix des Layers, W^3, besitzt die dieselben Dimensionen, $W^3 = [W_1^3; W_2^3; ...; W_3 2^3], \forall W_i^3 \in \mathbb{R}^{24\times24,n^2}$, mit

$$x^3 = \sigma(W^3 x^2). \tag{3}$$

Canonical View

Um die *Canonical View* aus den FIP zu erhalten, transformiert das *Fully Connected Layer* im vierten und letzten Block die Features aus x^3 entsprechend in ein zweidimensionales Bild in Graustufen. Die Gewichtsmatrix W^4 hat die Dimension $W^4 \in \mathbb{R}^{n_0,n_2}$, womit gilt

$$y = \sigma(W^4 x^3). \tag{4}$$

4 Training

Die Anzahl der zu berechnenden Gewichte stellt sich als unüberschaubar dar: Das neuronale Netzwerk ist zwar nicht überdurchschnittlich tief geschichtet, allerdings besitzen die Convolution Layers W^1, W^2, W^3 und W^4 viele individuelle Gewichte, wie in Kapitel 3 gezeigt wurde. Um das Training effizient zu gestalten, unterteilen Zhu et al. dieses nach dem Ansatz von Ngiam et al. in zwei Schritte: Initialisierung und Update [1, 8].

Die Gewichtsmatrizen der Pooling Layers, V^1 und V^2, lassen sich direkt nach Anweisung von Ngiam et al. initialisieren [8]. Die Gewichtsmatrizen der Convolution Layers hingegen, W^1, W^2, W^3 und W^4, sind zunächst nicht initialisiert und es wäre ineffizient, sie mit zufälligen Werten zu bestücken und eine schnelle Konvergenz während des Trainings zu erwarten [1]. Zunächst wird also versucht, eine Approximation der Gewichtsmatrizen mit der Minimierung des folgenden Ausgleichsproblems zu erreichen,

$$\underset{W^1,W^2,W^3,W^4}{\arg\min} \left\|\bar{y} - \sigma(W^4 x^3)\right\|_F^2 . \tag{5}$$

Die Berechnung von Formel 5 ist nicht trivial, da es sich um ein nichtlineares Problem handelt [1]. Da jedoch die Konstruktion des Netzwerks und der entsprechenden Convolution Layers nach den Tiled CNNs von Ngiam et al. geschieht, wissen wir, das W^1, W^2 und W^3 lineare Transformationen sind. Nun nähern wir deren Werte schrittweise über eine Reihe an linearen Ausgleichsproblemen an,

$$\underset{W^1}{\arg\min} \left\|\bar{y} - \sigma(OW^1 x^0)\right\|_F^2 , \tag{6}$$

$$\underset{W^2}{\arg\min} \left\|\bar{y} - \sigma(PW^2 x^1)\right\|_F^2 , \tag{7}$$

$$\underset{W^3}{\arg\min} \left\|\bar{y} - \sigma(QW^3 x^2)\right\|_F^2 , \tag{8}$$

$$\underset{W^4}{\arg\min} \left\|\bar{y} - \sigma(W^4 x^3)\right\|_F^2 . \tag{9}$$

Hier handelt es sich bei O, P und Q um wohldefinierte, binäre Matrizen, die die Werte der Feature Maps aus den jeweiligen x^i aufsummieren, um den berechneten Wert in dieselbe Dimension wie $\bar{y}$ zu transformieren. Betrachten wir das Netzwerk gilt es, den Fehler bei der Rekonstruktion des Gesichts zu minimieren,

$$E(X^0; W) = \|\bar{y} - y\|_F^2 . \tag{10}$$

Berechnen wir nach dem Backpropagation-Algorithmus die partielle Ableitung des Fehlers nach W^i, $\frac{\partial E}{\partial W_k^i}$, erhalten wir die Korrektur der Gewichte zu $W^i, i \in 1, 2, 3, 4$ zum k-ten Iterationsschritt via [1],

$$\Delta_{k+1} = 0,9 \cdot \Delta_k - 0,004 \cdot \varepsilon \cdot W_k^i - \varepsilon \cdot \frac{\partial E}{\partial W_k^i}, \tag{11}$$

$$W_{k+1}^i = \Delta_{k+1} + W_k^i. \tag{12}$$

Bei Δ_{k+1} handelt es sich um die vordefinierte *Momentum Variable* [13], ε bezeichnet die Lernrate [1]. Indem man nun die partielle Ableitung des Fehlers mit $\frac{\partial E}{\partial W_k^i} = x^{i-1}(e^i)^T$ schrittweise berechnet, lassen sich die Gewichtsänderungen, Δ_{k+1}, iterativ bestimmen [1].

5 Auswertung

Um das System zu testen und es in einen Vergleich zu etablierten Ansätzen wie LE [11], LGBP [14] und CRBM [15] setzen zu können, haben Zhu et al. mehrere Experimente durchgeführt [1]. Als Datensatz dazu werden Einträge aus der Gesichts-Datenbank MultiPIE verwendet, die insgesamt 756.204 Bilder der Gesichter von 337 verschiedenen Personen enthält [5]. Sie ist für diese Reihe an Experimenten geeignet, da das Gesicht jeder teilnehmenden Person von 15 Blickwinkeln und unter 20 verschiedenen Beleuchtungen fotografiert wurde [1]. Die Bilder aus MultiPIE wurden in vier verschiedenen Sitzungen aufgenommen [5]. Um einen vergleichbaren Datenbestand zu erhalten, wurde auf eine Untermenge der Einträge zugegriffen, die vollständig von allen Seiten ($-45°$ bis $45°$) fotografiert wurden. Dieser Bestand hat eine Größe von 128.940 Bildern und wurde ebenfalls von Autoren anderer Systeme, wie Asthana et al. und Li et al., verwendet [16, 7], womit eine Grundlage für den Vergleich von Systemen zur Erkennung von Gesichtern, insbesondere deren Merkmale, geschaffen wurde.

Die Experimente wurden in verschiedenen Phasen durchgeführt, *Setting-I*, *Setting-II* und *Setting-III*. Diese werden im Folgenden jeweils erläutert. Zum Training des Systems während jeder Phase werden die Bilder der jeweiligen Identitäten aus sieben Blickwinkeln ($-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, $45°$) verwendet.

Setting-I
: Für das erste Experiment der Serie, Setting-I, wurden lediglich Einträge aus MultiPIE mit neutraler Beleuchtung verwendet, um insbesondere die Stabilität des Ansatzes von Zhu et al. gegenüber verschiedener Blickwinkel zu testen. Von den insgesamt 337 Identitäten aus allen vier Sitzungen wurden Aufnahmen von 200 Personen zum Training des Systems verwendet, während die restlichen 137 Personen als Testeingaben dienten. Dafür wurden alle Gesichtsaufnahmen verwendet, ausgenommen Frontalaufnahmen mit einem Blickwinkel von 0° [1].

Setting-II
: Das zweite Experiment ist ähnlich wie Setting-I konzipiert, greift aber lediglich auf die Identitäten der ersten Sitzung von MultiPIE zurück. Dabei handelt es sich um 249 Stück, von denen 100 für das Training des Systems verwendet wurden. Die Aufnahmen der restlichen 149 Personen wurden als Stichproben zur Auswertung verwendet, ausgenommen derer mit einem Blickwinkel von 0° [1].

Setting-III
: Während des letzten Experiments, Setting-III, wird ebenfalls auf die Identitäten der ersten Sitzung aus MultiPIE zurückgegriffen. Um das System auch auf Vorhersagen unter verschiedenen Beleuchtungen zu testen, wählen Zhu et al. hier auch Datensätze aus, die nicht nur unter neutraler Beleuchtung entstanden sind.

6 Ergebnisse

Zhu et al. haben die in Kapitel 5 genannten Experimente durchgeführt und deren Ergebnisse präsentiert. Dazu haben sie die Erkennungsraten der FIP und der resultierenden rekonstruierten Ansichten mit denen aktueller Systeme verglichen, die Textur- und Learning-basierte Deskriptoren verwenden. Diese Ergebnisse werden in Kapitel 6.1 besprochen. Unter Verwendung der Canonical View lassen sich Varianzen in Beleuchtung und Blickwinkel neutralisieren, womit sich Ergebnisse anderer Ansätze verbessern lassen, wie in Kapitel 6.2 detailliert wird.

6.1 Vergleich mit anderen Ansätzen

Die Ergebnisse der Durchführung der Experimente unter Setting-I sind in Tabelle 1 abgebildet. Zum Vergleich der Erkennungsraten von FIP und Canoni-

Verfahren	Diskriminanzverf.	−45°	−30°	−15°	15°	30°	45°	Durchschn.
LGBP [14]		37,7	62,5	77,0	83,0	59,2	36,1	59,3
VAAM [16]		74,1	91,0	95,7	95,7	89,5	74,8	86,9
FA-EGFC [7]		84,7	95,0	99,3	99,0	92,9	85,2	92,7
SA-EGFC [7]		93,0	98,7	99,7	99,7	98,3	93,6	97,2
LE [11]	LDA	86,9	95,5	99,9	99,7	95,5	81,8	93,2
CRBM [15]	LDA	80,3	90,5	94,9	96,4	88,3	75,2	87,6
FIP	LDA	93,4	95,6	100,0	98,5	96,4	89,8	95,6
RL	LDA	95,6	98,5	100,0	99,3	98,5	97,8	98,3

Tabelle 1: Ergebnisse der Experimente dem ersten Experiment der Serie, *Setting-I* [1]. LGBP, VAAM und SA-EGFC müssen auf den jeweiligen Blickwinkel adjustiert werden, FA-EGFC, LE, CRBM, FIP und RL 'erkennen' ihn eigenständig. FIP bezeichnet die Face Identity-Preserving Features, RL die rekonstruierte Ansicht, die Canonical View.

cal View haben Zhu et al. auf Textur- sowie Learning-basierte Deskriptoren zurückgegriffen, deren Evaluierung ebenfalls mittels MultiPIE durchgeführt wurde: LGBP [14], VAAM [16], FA-EGFC [7], SA-EGFC [7], LE [11] und CRBM [15]. Auf einen Teil der Verfahren (LE, CRBM, FIP, RL) wurde eine lineare Diskriminanzanalyse (LDA) angewandt, um die Komplexität kalkulierten Features zu reduzieren [1]. Schließlich wurden die Ergebnisse aller Durchgänge pro Verfahren berechnet, mit den tatsächlichen Identitäten verglichen und sortiert nach Blickwinkel deren Durchschnitt berechnet. Die rechte Spalte von Tabelle 1 gibt den Gesamtdurchschnitt der Erkennungsrate des jeweiligen Systems an.

Wie der Gesamtdurchschnitt der Ergebnisse zeigt, ist die Fehlerquote der Erkennungsrate der Gesichter, die in der Canonical View rekonstruiert wurden, am geringsten. Diese bietet sogar passendere Vorhersagen als die Verfahren, die mit einem 3D-basierten Ansatz von Li et al. arbeiten (VAAM, FA-EGFC, SA-EGFC) [1]. Die Ergebnisse der Experimente Setting-II und Setting-III zeigen ähnliche Tendenzen, wie Zhu et al. berichten: In den folgenden Experimenten ist die Fehlerrate der mit RL bezeichneten rekonstruierten Ansichten deutlich geringer als die von Li et al. berichteten Fehlerraten [7]. Während Setting-II erreichte die Canonical View eine Erkennungsrate von 98,4%, während Setting-III unter verschiedenen Beleuchtungen allerdings nur eine Genauigkeit von 74,7% erzielt [1]. Zhu et al. stellen fest, dass diese niedrigere Genauigkeit an der Schwierigkeit von Vorhersagen unter verschiedenen Beleuchtungen liegt, aber dennoch im Vergleich besser als das Verfahren von Li et al. prognostiziert.

6.2 Erweiterung bestehender Ansätze

In Kapitel 2 wurde bereits diskutiert, dass Textur-basierte Deskriptoren zur Erkennung von Gesichtsmerkmalen (beispielsweise Gabor-Filter und Local Binary Patterns) diese zwar zuverlässig extrahieren, ihre Ergebnisse aber nicht stabil gegenüber Varianzen in Beleuchtung und Blickwinkel sind. Zusätzlich haben die von Zhu et al. berichteten Ergebnisse, die in Kapitel 6.1 detailliert wurden, gezeigt, dass selbst moderne Varianten dieser Deskriptoren keine stabile Erkennung garantieren können [1].

Die Daten der rekonstruierten Ansicht hingegen liefern überdurchschnittlich gute Erkennungsraten unter verschiedenen Beleuchtungen und Blickwinkeln, können Varianzen also weitestgehend neutralisieren und eine Zuordnung von Gesichtern zu den jeweiligen Personen ermöglichen. Da die Canonical View als zweidimensionales Bild vorliegt, sind sie wiederum als Eingabe für andere Methoden geeignet. Zhu et al. haben diese Daten genutzt, um die Rate positiver Ergebnisse von besagten Deskriptoren, unter anderem Gabor [6] und LBP [9], zu steigern: Unter den Parametern von Setting-I (beschrieben in Kapitel 5) wurden die verschiedenen Deskriptoren auf die jeweiligen Bilder mit und ohne Neutralisierung durch Canonical View angewandt und die jeweiligen Erkennungsraten analysiert.

Die invarianten Eigenschaften der Canonical View zeigen sich deutlich in den Ergebnissen. Während Gabor und LBP bei neutralem Blickwinkel ($-15°$ bis $15°$) Erkennungsraten von 90% bis 100%, liegen sie bei starker Abweichung ($-45°$ und $45°$) deutlich unter 70% [1]. Wenden Zhu et al. hingegen die Deskriptoren auf die zuvor berechnete Canonical View der Eingabe an, erzielen die Ansätze unter allen Blickwinkeln Ergebnisse zwischen 80% und 100% [1].

Somit zeigen die Autoren, dass ihr Ansatz nicht nur vielseitige Gesichtsmerkmale extrahieren kann, die eine erfolgreiche Zuordnung von Gesichtern von Personen ermöglichen. Zusätzlich kann eine rekonstruierte, neutrale Frontalansicht auf Basis dieser Merkmale dazu verwendet werden, Erkennungsraten bereits existierende Systeme deutlich zu steigern.

7 Zusammenfassung

Zhu et al. haben ein künstliches, vielschichtiges neuronales Netzwerk vorgestellt, das Gesichter zuverlässig und unabhängig von Kopfneigung und Beleuchtung anhand von Face Identity-Preserving Features beschreiben kann. Zusätzlich stellen sie mit der Canonical View eine neutrale, rekonstruierte Darstellung dieses Gesichts in Form einer Frontalansicht vor. Die Verwendung vieler Eigenheiten von Convolutional Neural Networks erhöht die Stabilität des Netzwerks und ermöglicht effizientes Training der Gewichte. Ein Vergleich hat erwiesen, dass das von Zhu et al. konzipierte System zuverlässiger als andere zeitgenössische Ansätze zur Gesichtserkennung arbeitet und es ermöglicht, Gesichter den jewei-

ligen Personen erfolgreich zuzuordnen. Des Weiteren geben die Autoren einen Ausblick darauf, wie mithilfe der Canonical View etablierte Systeme deutlich positivere Ergebnisse erzielen können.

Literatur

[1] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 113–120, 2013.

[2] Ashok Samal and Prasana A Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition*, 25(1):65–77, 1992.

[3] About face id advanced technology. https://support.apple.com/en-us/HT208108, 2017. URL `https://support.apple.com/en-us/HT208108`. [Online; accessed 17-January-2018].

[4] The cmu multi-pie face database. http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Content.html, 2010. URL `http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Content.html`. [Online; accessed 24-January-2018].

[5] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[6] Laurenz Wiskott, Norbert Krüger, N Kuiger, and Christoph Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):775–779, 1997.

[7] Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan. Morphable displacement field based image matching for face recognition across pose. *Computer Vision–ECCV 2012*, pages 102–115, 2012.

[8] Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pang W Koh, Quoc V Le, and Andrew Y Ng. Tiled convolutional neural networks. In *Advances in neural information processing systems*, pages 1279–1287, 2010.

[9] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[10] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1): 51–59, 1996.

[11] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2707–2714. IEEE, 2010.

[12] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[13] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[14] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 786–791. IEEE, 2005.

[15] Gary B Huang, Honglak Lee, and Erik Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. IEEE, 2012.

[16] Akshay Asthana, Tim K Marks, Michael J Jones, Kinh H Tieu, and MV Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 937–944. IEEE, 2011.