Westfälische Wilhelms-Universität Münster

Praktikum zur Mustererkennung

Emotionserkennung

Klassifikation von Action Units anhand von Landmarks

Robin Rexeisen
Matrikelnr. 123456

Johannes Stricker
johannesstricker@gmx.net
Matrikelnr. 383779

Alexander Schlüter
alx.schlueter@gmail.com
Matrikelnr. 409649

Betreuer:
Sören Klemm
soeren.klemm@wwu.de

eingereicht am 16. September 2016

Fachbereich 10
Mathematik und
Informatik

Vorwort

Hier entsteht ein Vorwort.

Inhaltsverzeichnis

1 Einleitung

TODO

1. Einführung in das Problems
2. Was für Daten wir zur verfügung (float-array) haben, woher die kommen (link auf Paper o.ä.).
3. Aufbau der Datenbank (mehrere Personen, ein Länge des Videos, ...)
4. Action-Units beschreiben (nur die in den Daten vorkommen)

2 Methodik

Hier keine Details zur Implementierung!

2.1 Vorverarbeitung

2.1.1 Punktwolken

1. Beschreibung der Daten ?!
2. Normalisierung
3. Randomisiertes erweitern
4. PCA

2.1.2 Feature Extrahieren

Statische Features

Dazu schreiben, wieso wir finden, dass das Feature ein gutes ist (z.B. weil es den Feature-Raum verkleinert, ...)

1. XYFeature
2. Orientation
3. EuclidianDistance
4. CenterDistance
5. CenterOrienation
6. Interpolation

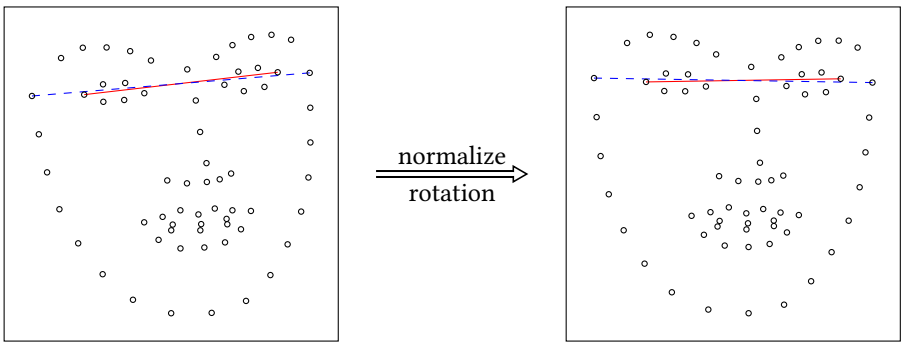

Abbildung 2.1: Normalisierung der Rotation anhand ausgesuchter Linien

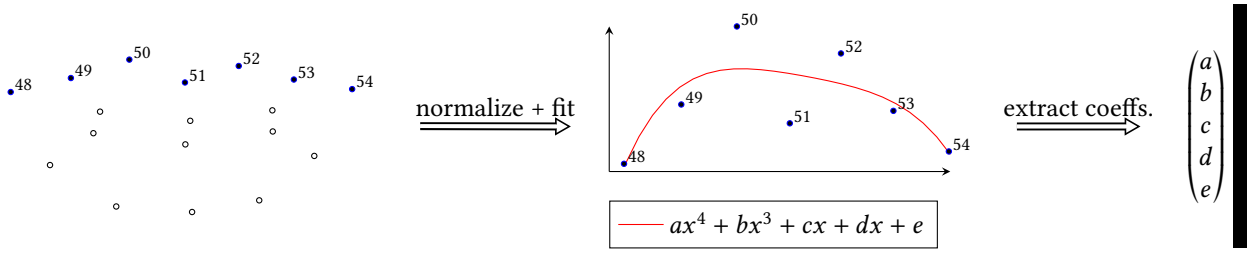

Abbildung 2.2: InterpolationFeatureExtraction

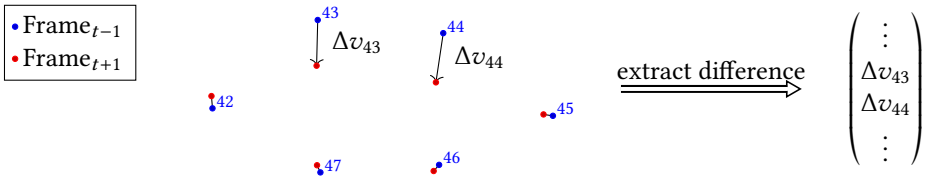

Abbildung 2.3: TimeDifferentialExtraction

Zeitliche Features

- TimeDifferential

Featureverarbeitung

Zweck mitbeschreiben (z.B. PCA -> FeatureRaum weiter reduzieren)

- Negativanteil verringern
- MinMax/MeanVar Normalisieren
- Shufflen
- PCA

Klassifikatoren

SVM + Random Forests, Art von Parametern
Vielleicht noch eine Beschreibung einer allgemeinen Pipeline.

2.2 Evaluierungsmethoden

Die im vorherigen Abschnitt beschrieben Methoden zur Feature Extraction, Verarbeitung und Klassifikation sollen in verschiedenen Kombinationen evaluiert werden. Der erste Datensatz aus 10 Personen wird dazu aufgeteilt in 60% Trainingsmenge und 40% Validierungsmenge. Hier ist die Entscheidung zu treffen, wie die Personen auf die Mengen aufgeteilt werden:

1. Erst die Frames durchmischen, dann aufteilen: Dies ist sinnvoll, wenn der Klassifikator nur verwendet werden soll, um Action Units in neuen Frames von schon bekannten Personen zu erkennen. Es wird nicht getestet, wie gut der Klassifikator auf neue Personen generalisiert!

2. 6 Personen nur im Training, 4 nur in der Validierung verwenden: Die Performance auf der Validierungsmenge ist repräsentativ dafür, wie gut der Klassifikator Action Units bei bisher unbekannten Personen erkennt

Erste Tests haben gezeigt, dass Methode 1 zu deutlich besserern Performancestatistiken führt. Wir haben uns aber für Methode 2 entschieden, weil die Generalisierung auf neue Personen das interessantere Problem ist: In Anwendungsfällen ist es wünschenswert, für neue Personen nicht erst mehrere tausend Frames manuell labeln zu müssen, um den Klassifikator auf dieser Person zu trainieren.

Aufgrund der geringen Anzahl positiver Samples (Frames, in denen die Action Unit aktiviert ist), ist die Accuracy keine zuverlässige Statistik. Ein Klassifikator, der die Action Unit immer als "nicht aktiv" klassifiziert, könnte sehr hohe Accuracy erreicht, ohne tatsächlich etwas über die Action Unit gelernt zu haben. Stattdessen evaluieren wir die Klassifikatoren anhand von

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \qquad \text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Der F1 score ist das harmonische Mittel zwischen Precision und Recall. Da wir von der Aufgabenstellung her keine Präferenz für hohe Precision / hohen Recall haben, nutzen wir den F1 score als erste Zahl zum Vergleich der Klassifikatoren.

Es werden alle Kombinationen aus Feature Extraction, Klassifikator und Parametern auf dem ersten Datensatz trainiert und evaluiert. Anschließend werden die besten fünf pro Action Unit anhand des F1 scores ausgewählt. Da pro Action Unit ca. 160 Kombinationen evaluiert werden, kann es durch diese Auswahl der besten fünf zu einem Overfitting gegen die Validierungsmenge kommen. Um realistische Zahlen für die Performance zu bekommen, werden deshalb die besten fünf nochmal auf einer Testmenge evaluiert. Diese besteht aus fünf bisher unbekannten Personen aus einem zweiten Datensatz.

- Welche Parameter/Pipeline zum trainieren
  - Warum diese Parameter und keine anderen?

3 Implementierung

- Architektur/OS eines lauffähigen Systems
- Softwareabhängigkeiten
- Programmiersprache

3.1 QViewer

1. Zweck
2. Bilder

3.2 Auto-Train

- Zum trainieren und evaluieren
- Kurzes Wort zum Design von FeatureExtractor
- Automatisches Speichern aller relevanten Dateien.
- Erwähnung der JSON-Konfigurations-Datei
  - Design von Processors

4 Ergebnis

AU	Bester Klassifikator					
	F1 Val	F1 Test	Prec. Test	Recall Test	Klassifikator	Features
Lips Part	0.656	0.575	0.598	0.554	RF	CenterDist.
Upper Lid Raiser	0.394	–	0	0	SVM Polyn.	Interpolation
Outer Brow Raiser	0.391	–	0	0	SVM Polyn.	EuclidianDist.
Lip Corner Puller	0.37	0.431	0.317	0.674	SVM Lin.	XY
Cheek Raiser	0.277	0.18	0.102	0.769	SVM Polyn.	XY
Nose Wrinkler	0.242	0.03	0.015	0.62	SVM Polyn.	EuclidianDist.
Inner Brow Raiser	0.2	0.084	0.064	0.125	SVM Polyn.	EuclidianDist.
Chin Raiser	0.191	0.03	0.163	0.017	SVM Polyn.	CenterDist.
Brow Lowerer	0.163	0.2	0.659	0.118	RF	EuclidianDist.
Lip Corner Depressor	0.022	–	0	0	SVM Polyn.	XY

Tabelle 4.1: F1 scores und Testergebnisse des besten Klassifikators pro Action Unit

In Tabelle 4.1 sind die Ergebnisse des besten Klassifikators (ausgewählt nach F1 score auf der Validierungsmenge) pro Action Unit zu sehen. Gute Klassifikation auf unbekannten Personen ist möglich für Lips Part: Der hohe F1 score 0.656 in der Validierung bestätigt sich auch auf der Testmenge. Akzeptable Performance liefert der beste Klassifikator für Lip Corner Puller. Dieser verbessert sich sogar von einem F1 score von 0.37 in der Validierung auf 0.431 im Test.

Die Klassifikatoren für Outer Brow Raiser und Upper Lid Raiser scheinen in der Validierung akzeptabel, erkennen jedoch im Test überhaupt keine Aktivierung der Action Units mehr. Die übrigen Action Units werden mit keiner unserer Feature Extraction-Methoden an neuen Personen befriedigend klassifiziert.

Abb. 4.1 zeigt, dass die verschiedenen Kombinationen aus Feature Extraction, Klassifikator und Parametern zu stark variierender Performance auf der Validierungsmenge führen. Der Tradeoff zwischen Precision und Recall ist deutlich zu sehen. Man sieht eine Trennung zwischen SVM und Random Forest Klassifikatoren: erstere tendieren dazu, zu viele negative Frames (ohne Aktivierung der Action Unit) als positiv zu klassifizieren, was zu schlechter Precision führt. Die Random Forests neigen hingegen zu vielen False Negatives. Sie schneiden bezogen auf Lips Part besser ab, bei anderen Action Units ist die Performance zwischen den Klassifikatoren ausgeglichen.

F1 Val	F1 Test	Prec. Test	Recall Test	Klassif.	Parameter	Features
0.656	0.575	0.598	0.554	RF	#trees = 20, maxDepth = 20	CenterDist.
0.656	0.594	0.639	0.555	RF	#trees = 50, maxDepth = 20	CenterDist.
0.647	0.655	0.809	0.550	RF	#trees = 20, maxDepth = 10	CenterDist.
0.623	0.616	0.966	0.452	RF	#trees = 50, maxDepth = 10	Interpolation
0.585	0.732	0.752	0.713	RF	#trees = 20, maxDepth = 20	XY

Tabelle 4.2: F1 scores und Testergebnisse der Top 5 Klassifikatoren für Lips Part

Die Dominanz der Random Forests für Lips Part ist auch in Tabelle 4.3 zu sehen. Die beste SVM taucht mit einem F1 score von 0.356 in den Top 5 nicht mehr auf.

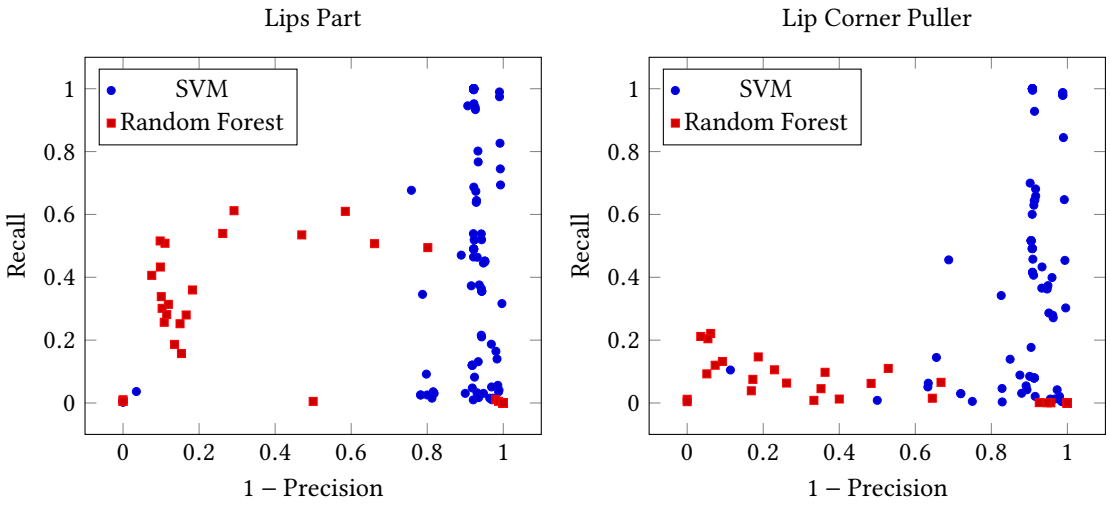

Abbildung 4.1: Validierungsergebnisse für Lips Part und Lip Corner Puller. Jeder Punkt steht für eine Kombination aus Feature Extraction, Klassifikator und Parametern.

F1 Val	F1 Test	Prec. Test	Recall Test	Klassif.	Parameter	Features
0.370	0.431	0.317	0.674	SVM Lin.	–	XY
0.358	0.514	0.675	0.415	RF	#trees = 10, maxDepth = 4	Interpolation
0.347	0.389	0.752	0.262	RF	#trees = 50, maxDepth = 20	Interpolation
0.336	0.536	0.733	0.423	RF	#trees = 20, maxDepth = 10	Interpolation
0.248	0.426	0.774	0.293	RF	#trees = 20, maxDepth = 10	EuclidianDist.

Tabelle 4.3: F1 scores und Testergebnisse der Top 5 Klassifikatoren für Lip Corner Puller

Features	Bester F1 score	Dimensionen	Dim. nach PCA
CenterDistance	0.656	66	17
Interpolation	0.623	37	–
XY	0.585	132	28
EuclidianDistance	0.5	2,145	30
Orientation	0.144	2,145	2,145
CenterOrientation	0.137	66	2
TimeDiff_XY	0.126	132	117
TimeDiff_Interpolation	0.06	37	–

Tabelle 4.4: Vergleich der verschiedenen Features für Lips Part. Auf Interpolation-Features wurde keine PCA angewandt.

5 Diskussion

- Wieso sind gut bei Test, aber schlecht bei Training
- Wieso sind diese Klassifikatoren gut und andere nicht
- Wieso sind viele Action-Units schlecht zu erkennen
- Wieso gerade diese Feature so gut?
- ⇒ Overfitting
- Warum geht diese Action-Unit besser als andere

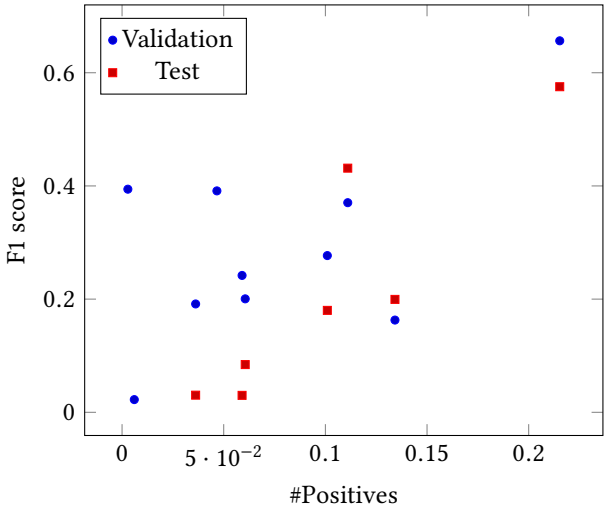

Abbildung 5.1: F1 score des besten Klassifikators jeder Action Unit gegen Anteil positiver Samples

6 Fazit

6.1 Zusammenfassung

6.2 Ausblick

- Was könnte man noch verbesser, und wieso haben wir das nicht gemacht (z.B. aus Zeitgründe)
    1. Mehr Kombinationen (Mit/ohne PCA, mehr Time-Differential-Feature, überhaupt mehr zeitliche Features, andere normalisierungen der Punktwolke, Neuronales-Netzwerk oder andere Klassifikatoren dazu benutzen)
- Wie könnte das Ergebnis besser werden (z.B. mehr Daten von mehreren Personen)