Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Computer Vision for Music Identification

Yan Ke, Derek Hoiem,

Rahul Sukthankar

Inhalt

- Ziel und Anforderungen
- Problemstellen
- Ansatz im Überblick
- Ansatz im Detail
- Resultate
- Exkurs Talkalyzer

Ziel und Anforderungen

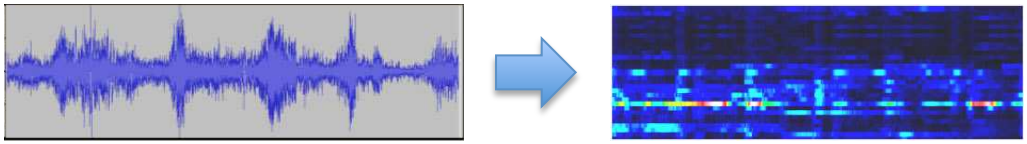
- Ziel der Arbeit
  - Identifizierung von Musikstücken anhand weniger Sekunden gestörter Audiodaten
- Anforderungen an das System
  - Hohe Ausbeute (Recall)
  - Hohe Präzision (Precision)
  - Schnelles Retrieval

Problemstellen

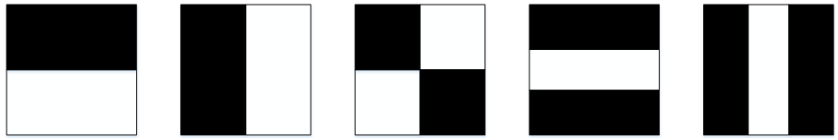
- Queries
  - Schlechte Aufnahmequalität
    - resistent gegen Störgeräusche und schlechte Aufnahmequalität
    - Ausschluss von stark gestörten Abschnitten
  - Aufnahme an einem willkürlichen Punkt im Song
    - lokal und robust gegenüber kleinen zeitlichen Verschiebungen
- Datenmenge
  - DB mit mehreren 100’000 Songs ist zu erwarten
    - effizientes berechnen und indexieren
    - effizientes Abfragen

Ansatz im Überblick

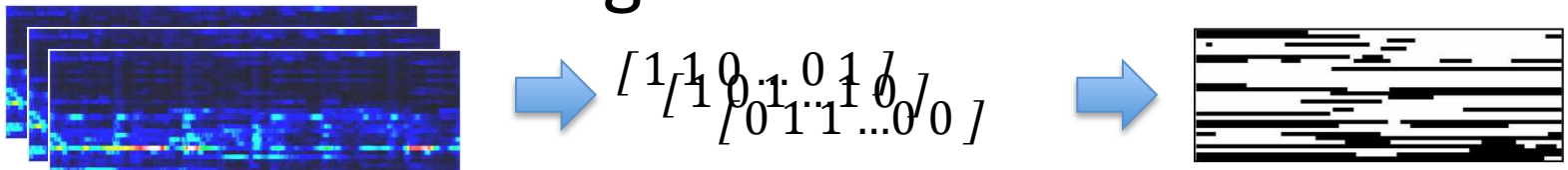
- Darstellung als Spektrogram



- Lernen eines Filtersets



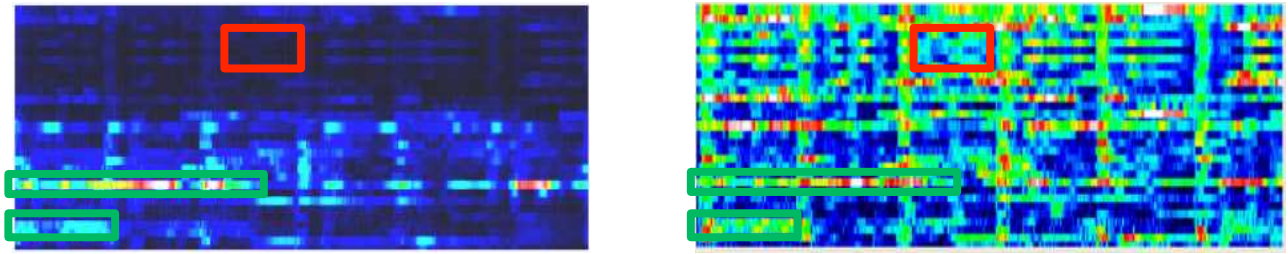
- Extrahieren der Signatur



- Retrieval

Darstellung als Spektrogram

- Spektrogramme zeigen Ähnlichkeit/Differenz von Original und Aufnahme

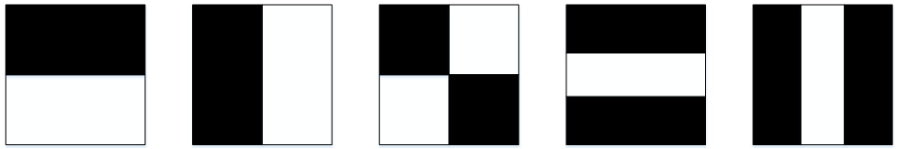


- Transformation mittels *Short-Term Fourier Transform* (STFT)
  - 33 logarithmisch verteilte Frequenzbändern
  - gemessen über Zeitfenster von 0.372 s
  - jeweils um 11.6 ms inkrementiert
  - Rund 97% Überlappung

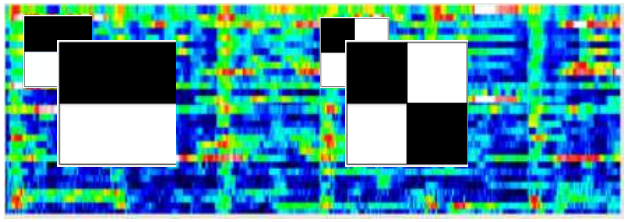
Lernen eines Filtersets (1) Filterkandidaten

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- "Einfaches" Vergleichen von Spektrogrammen ungenau und langsam
- Set von visuellen Filtern
  - Anforderungen
    - robust gegenüber kleinen zeitlichen Verschiebungen
    - halten Ähnlichkeit/Differenz fest
    - ignorieren irrelevante Daten
  - 5 Basisfilter
    - von Viola und Jones zur automatischen Erkennung von Objekten entworfen



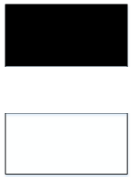
  - Variationen in exponentiellen Schritten
    - Rund 25'000 Filterkandidaten bzw. *weak Classifiers*.



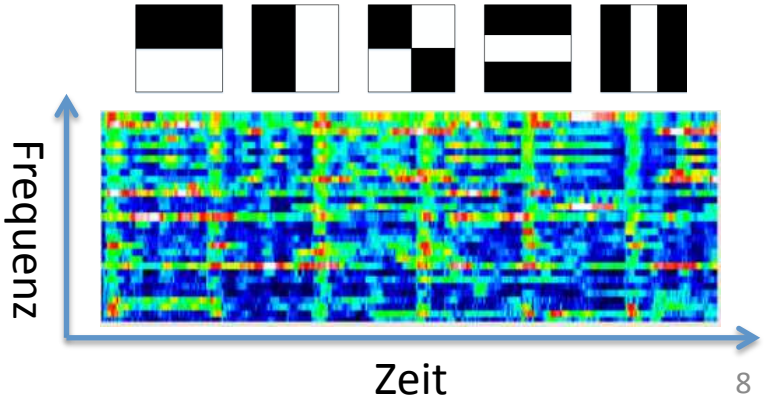
Lernen eines Filtersets (2) Filtereigenschaften

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Min. 1 schwarzes und min. 1 weisses Rechteck
- Funktion
  - Aufsummieren der Pixelwerte in der schwarzen Fläche
  - Aufsummieren der Pixelwerte in der weissen Fläche
  - Flächen subtrahieren → et voilà



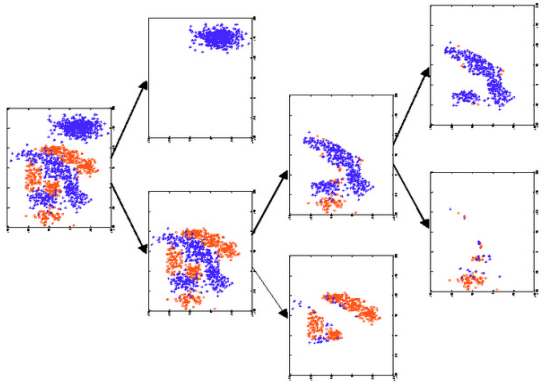
- Filter "besitzen" einen Threshold
  - Filterwert oberhalb → 1
  - Filterwert unterhalb → 0
- Interpretation
  - 1. Energiedifferenzen in benachbarten Frequenzbändern
  - 2. Energiedifferenzen über die Zeit
  - 3. Verschiebungen der dominanten Frequenz über die Zeit
  - 4. Energiespitzen in benachbarten Frequenzbändern
  - 5. Energiespitzen über die Zeit



Lernen eines Filtersets (3) AdaBoost

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Selektion von *n* Filtern mit AdaBoost
  - Berechnen eines «Strong Classifiers»
    - H(x1, x2) → y = {−1, 1}
    - besteht aus *n* «weak classifiers»
  - «weak classifier»
    - hm(x1, x2) = sgn[(fm(x1) − tm)(fm(x2) − tm)]
    - Hier: Viola-Jones Filter inkl. Threshold
    - nur besser als **zufällig** Songs unterscheiden/erkennen
  - Wählt iterativ die «besten» Filter
    - Trainigsdaten sind gewichtet
      - Hohes Gewicht bei falsch erkannten Daten
      - Tiefes Gewicht bei korrekt erkannten Daten
    - 1 Filter pro Iteration
    - Anpassung **aller** Gewichte pro Iteration

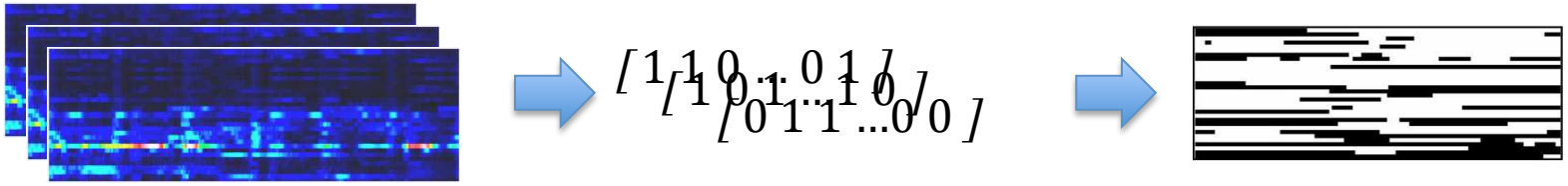


Lernen eines Filtersets (4) Pairwise Boost

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Trainingsdaten
  - Positive Paare (Original und Aufnahme des selben Songs)
  - Negative Paare (verschiedene Songs)
- Problem AdaBoost
  - «weak classifiers» **nicht besser als Zufall** bei negativen Paaren
  - verletzt «weak classifier» Bedingung von Adaboost
- Pairwise Boost ist eine Anpassung von AdaBoost
  - Nur positive Paare werden neu gewichtet
  - Gewichtung der positiven und negativen Paaren auf jew. 0.5 normalisiert
- Experimente zeigten:
  - n = 32 als geeignete Wahl
  - “lose korrelierten” Daten brauchen keine negativen Paare
    - AdaBoost evtl. doch geeignet!?

Extrahieren der Signatur



- Die Filter (inkl. Thresholds) liefern einen Desktriptor pro Spektrogramm
  - 32-Bit Vektor
  - Kompakte Repräsentation einer **lokalen** Region (eines Songs)
  - Enthalten nicht genügend Information um zuverlässig Songs zu erkennen
- Signaturen sind kombinierte, überlappende Deskriptoren
  - Spektrogramme:
    - Rund 97% Überschneidung
- Signatur als Basis für das Matching und Retrieval

Retrieval (1) Datenbank

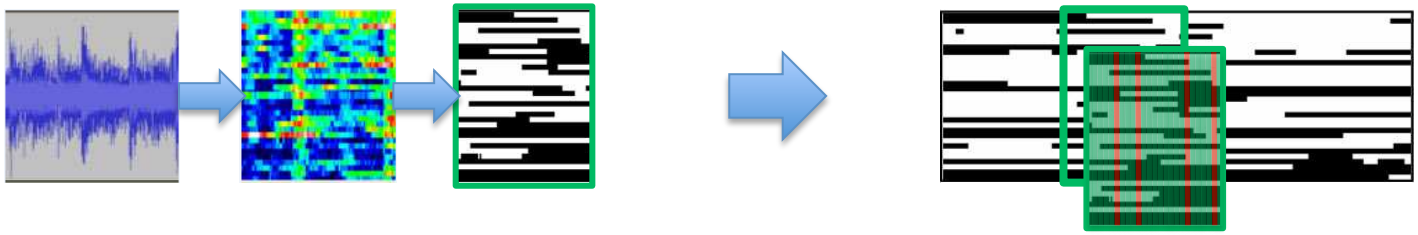
- Deskriptoren erlauben direktes indexieren
  - Verwendung von Hashtables
    - Sehr schnell
    - kaum Verlust der Genauigkeit
  - Abfrage der Signaturen anhand der Deskriptoren

Retrieval (2) Matching

- Finden der “ähnlichen” Signaturen
  - Hamming-Distanz
  - Deskriptoren mit einer Hamming-Distanz <= 2
    - gelten als “near-neighbors” und sind somit relevant
- Selektion des besten Matchs
  - Zeitliches Ausrichten der Query
    - RANSAC-Algorithmus (random sample consensus)
  - Übereinstimmung berechnen
    - Expectation-Maximization (EM) Score als Distanz
- Vergleich mit Threshold
  - EM Score > Threshold = Treffer

Retrieval (3) Ausschluss

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



- Einzelne Deskriptoren können von «Störgeräuschen» dominiert sein
  - diese sollen (mögl.) ignoriert werden
- Ausschlussmodell
  - Gegeben: Signatur $x\uparrow r = (x\downarrow 1\uparrow r,\ x\downarrow 2\uparrow r, ..., x\downarrow n\uparrow r)$
  - Vergleich mit Signatur $x\uparrow o = (x\downarrow 1\uparrow\uparrow o, x\downarrow 2\uparrow o, ..., x\downarrow n\uparrow o)$
  - $x\downarrow i\uparrow r-o$ = Bit-Differenz zwischen $x\downarrow i\uparrow r$ und $x\downarrow i\uparrow o$
  - Verteilung von $x\downarrow i\uparrow r-o$ mittels unabhängigen Bernoulli-Zufallsvariablen modellieren
  - Mit EM und einem entsprechenden Threshold Label $y\downarrow i$ bestimmen
    - $y\downarrow i=1$ wenn vom Musikstück generiert
    - $y\downarrow i=0$ wenn von Störungen überlagert

Experimente (1)
Setup

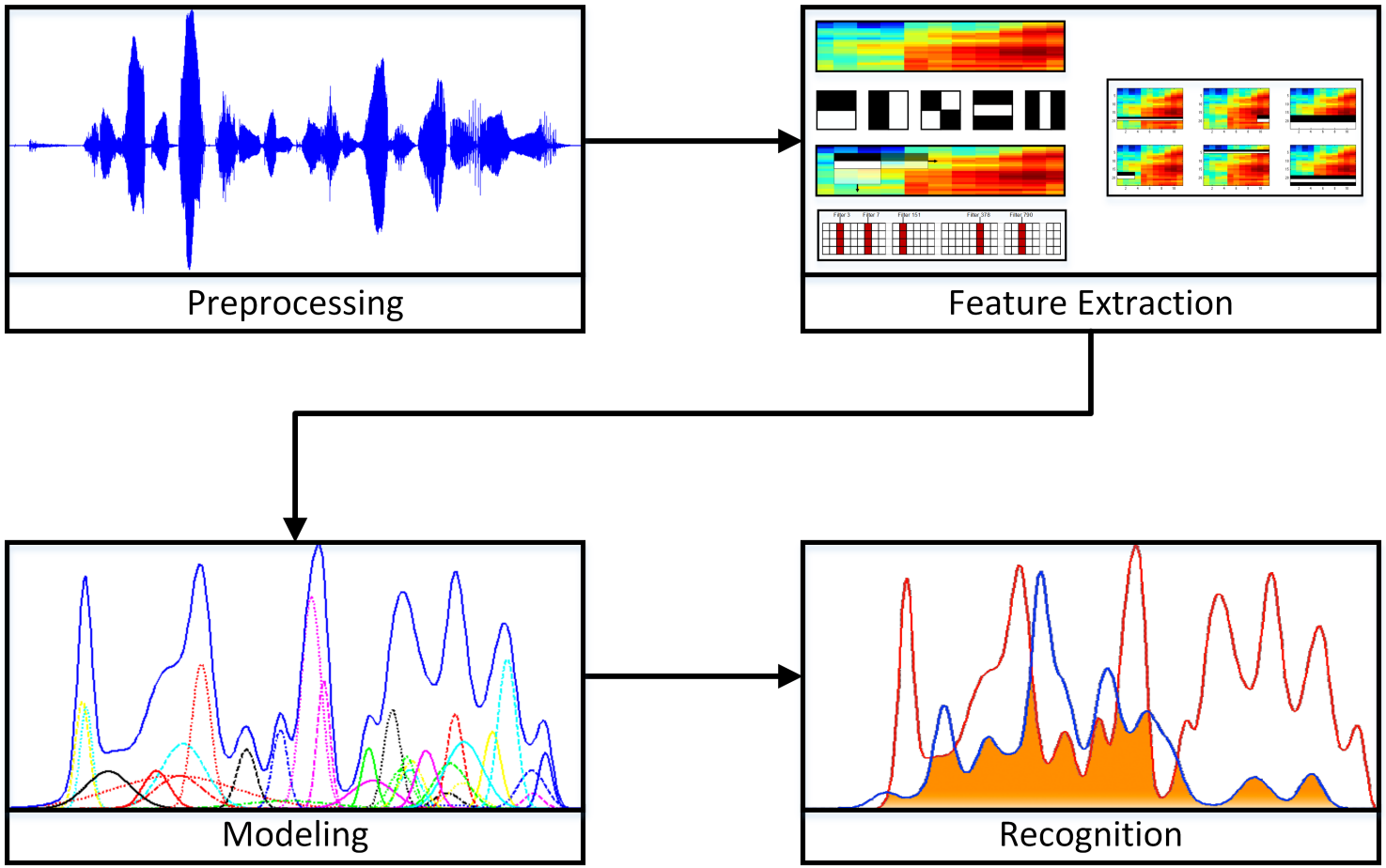
- Total 1862 Musikstücke
  - Grosse Auswahl an Genres enthalten
- Trainingdata
  - 78 Musikstücke
  - Abgespielt mit low-quality Speakers
  - Aufgenommen mit low-quality Mikrophonen
- Test A
  - 71 Musikstücke
  - Abgespielt auf niedriger Lautstärke
  - Aufgenommen mit einem verzerrenden Mikrophon
- Test B
  - 220 Musikstücke
  - Aufgenommen in einer sehr lärmigen Umgebung

Experimente (2)
Resultate

- Gegeben: 10 Sekunden Daten pro Query
- Test A
  - Recall: 90%
  - Precision: 96%
- Test B
  - Recall 80%
  - Precision 93%

Exkurs Talkalyzer (1) Überblick

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



Exkurs Talkalyzer (2) Resultate

- Setting:
  - 40 Sprecher aufgeteilt in 80 Gruppen
  - Clustering experiment
  - Baselineansatz: MFCC / GMM
- Errorrate:
  - Baseline: 12.76%
  - Talkalyzer: 16.95%
- Erkenntnis:
  - Problemstellung zu unterschiedlich bzw. Anpassung an Sprecherclustering zu simpel

Fragen