INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

AUDIO LABS

Weakly Supervised Multi-Pitch Estimation Using Cross-Version Alignment

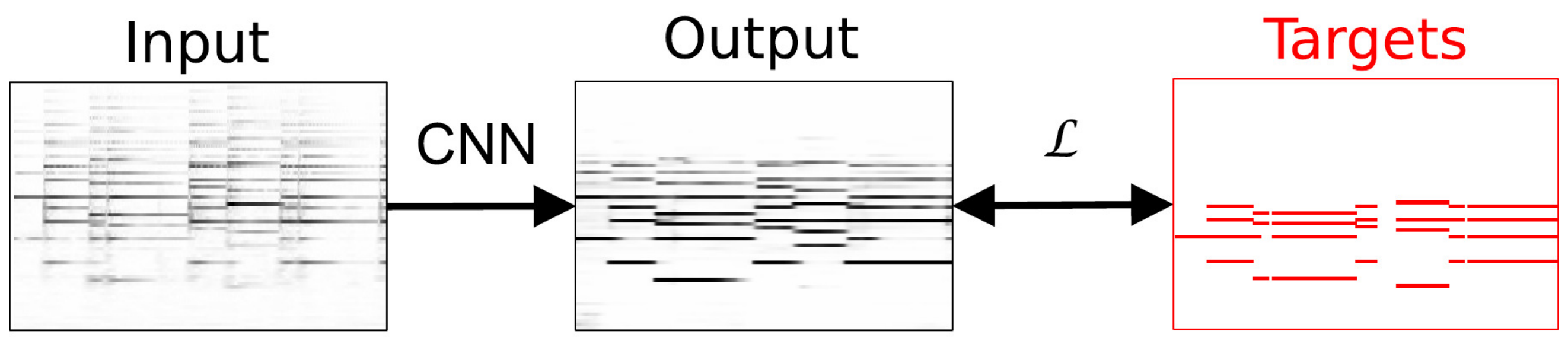
Michael Krause, Sebastian Strahl, Meinard Müller

Abstract

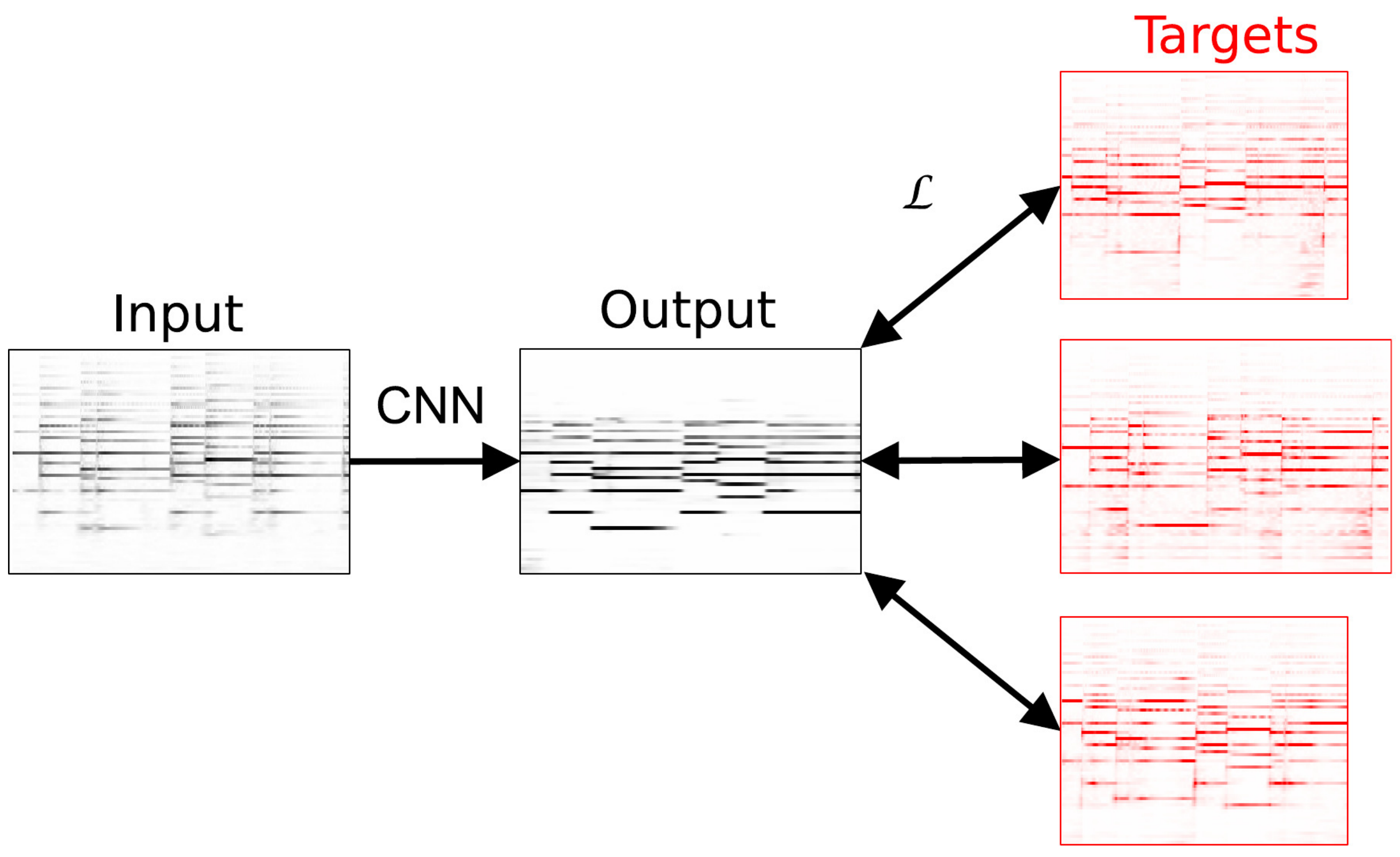
Multi-pitch estimation (MPE), the task of detecting active pitches within a polyphonic music recording, has garnered significant research interest in recent years. Most state-of-the-art approaches for MPE are based on deep networks trained using pitch annotations as targets. The success of current methods is therefore limited by the difficulty of obtaining large amounts of accurate annotations. In this contribution, we propose a novel technique for learning MPE without any pitch annotations at all. Our approach exploits multiple recorded versions of a musical piece as surrogate targets. Given one version of a piece as input, we train a network to minimize the distance between its output and time–frequency representations of other versions of that piece. Since all versions are based on the same musical score, we hypothesize that the learned output corresponds to pitch estimates. To further ensure that this hypothesis holds, we incorporate domain knowledge about overtones and noise levels into the network. Overall, our method replaces strong pitch annotations with weaker and easier-to-obtain cross-version targets.

1. Introduction

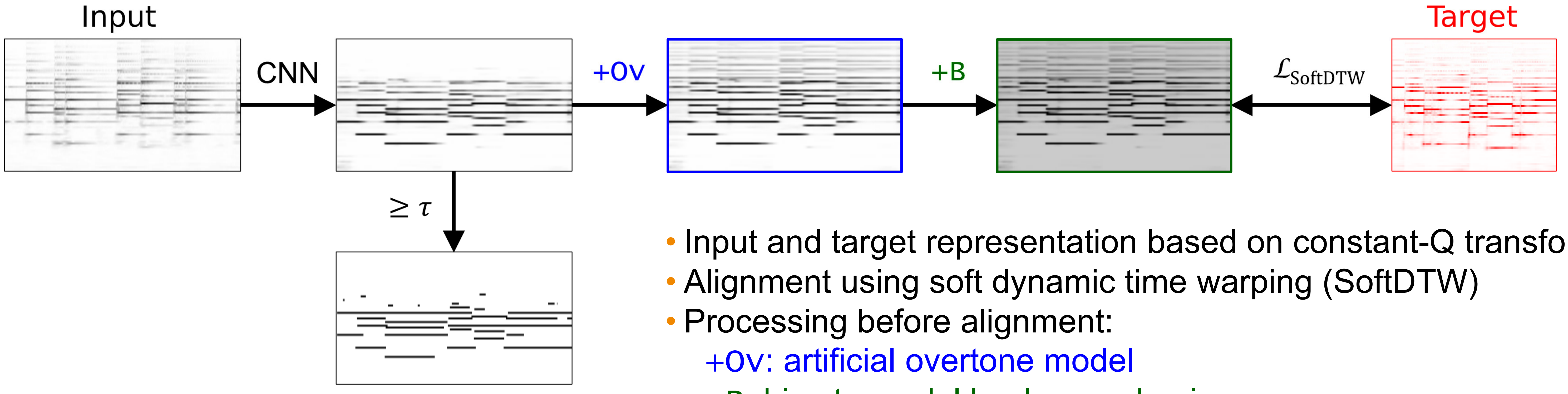
- **Task:** multi-pitch estimation (MPE)
- **Usual approach:** fully supervised
- **Problem:** needs pitch annotations as targets



- **Idea:** learn MPE by finding commonalities between different versions → **weakly supervised**
- Learn from cross-version correspondences



2. Proposed Approach



- Input and target representation based on constant-Q transform (CQT)
- Alignment using soft dynamic time warping (SoftDTW)
- Processing before alignment:
  - +Ov: artificial overtone model
  - +B: bias to model background noise
- Apply threshold τ for final MPE output

3. Experiments

- Dataset: Schubert Winterreise
  - singing & piano
  - 24 songs, multiple versions per song
  - challenging train-test split
- Baseline `CQT`: directly threshold CQT
- Upper bound `Sup`: fully supervised, pitch annotations

Scenario	AP	F-measure
CQT	0.410	0.443
Ours	0.589	0.585
Ours+Ov	0.639	0.553
Ours+B	0.563	0.560
Ours+Ov+B	0.646	0.625
Sup	0.753	0.700

AP = Average Precision

4. Summary

- Learn **MPE without pitch annotations**
- Use **different versions** of same music piece as **surrogate targets**
- Usable multi-pitch estimates, but still performance gap to fully supervised approach

- Future work:
  - larger models & datasets
  - more advanced, DDSP-like synthesis

References

[1] Marco Cuturi and Mathieu Blondel, “Soft-DTW: a differentiable loss function for time-series,” in ICML, 2017.
[2] Michael Krause, Christof Weiß, and Meinard Müller, “Soft dynamic time warping for multi-pitch estimation and beyond,” in ICASSP, 2023.

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

Fraunhofer IIS