ISMIR MTL2020

Dance Beat Tracking from Visual Information Alone

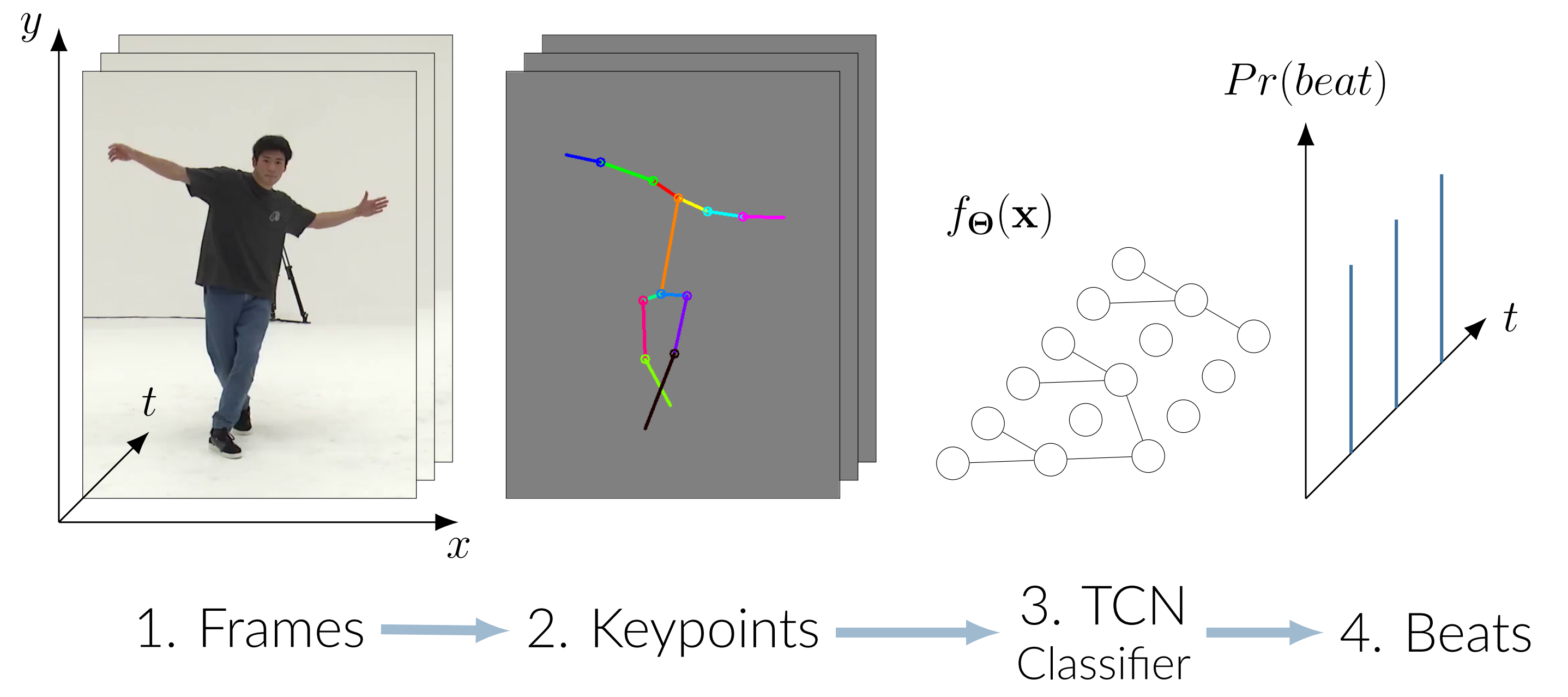
Fabrizio Pedersoli and Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST), Japan

1. Introduction

- Dance Information Retrieval (DIR)
  - Extracting *high-level* semantics information from *dance videos*
  - DIR tasks similar to Music Information Retrieval (MIR)
  - DIR tasks typically solved by analyzing the *visual information*
- Dance Beat Tracking (without audio signals)
  - Unexplored fundamental topic in DIR research
  - Detection of musical beats by using visual information
  - *Classify* each video frame as "beat" or "non-beat" frame
- Important **applications** of dance beat tracking
  - Automatic *synchronization* of dancing with music
  - Temporal *alignment* of videos (time stretching)
  - Identification of *out-of-sync* dance videos

2. Approach



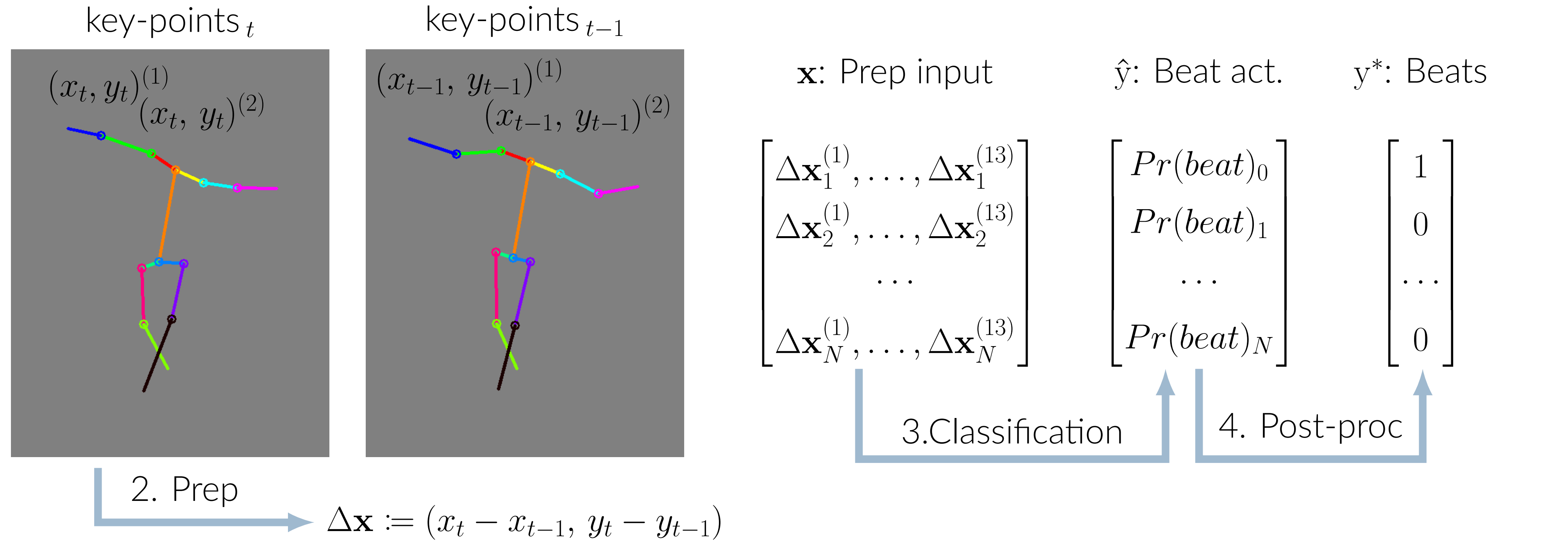
Step 1. **Body key-points** are extracted from video frames by OpenPose

Step 2. Body key-points are **pre-processed**

Step 3. Sequence of pre-processed key-points is classified by a Temporal Convolutional Neural Network (**TCN**) (output is the beat activation function)

Step 4. Beat activation is **post-processed** to get the final beats positions

3. Classification



$$\begin{bmatrix} \Delta\mathbf{x}_1^{(1)}, \dots, \Delta\mathbf{x}_1^{(13)} \\ \Delta\mathbf{x}_2^{(1)}, \dots, \Delta\mathbf{x}_2^{(13)} \\ \dots \\ \Delta\mathbf{x}_N^{(1)}, \dots, \Delta\mathbf{x}_N^{(13)} \end{bmatrix} \quad \begin{bmatrix} Pr(beat)_0 \\ Pr(beat)_1 \\ \dots \\ Pr(beat)_N \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}$$

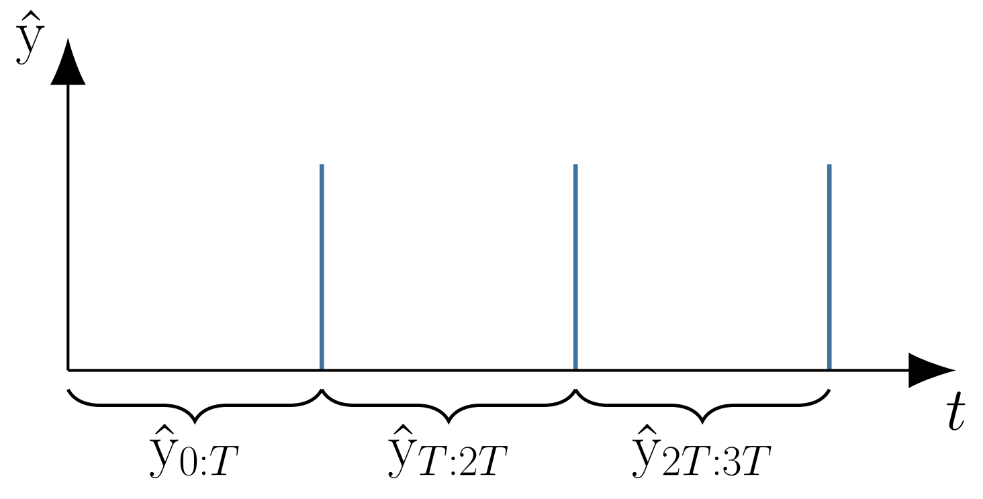
$$\Delta\mathbf{x} \coloneqq (x_t - x_{t-1},\, y_t - y_{t-1})$$

- Step 2: **Pre-process**
  - Sequence of (x, y) *absolute coordinates* of body key-points is converted into frame-by-frame $(\Delta x, \Delta y)$ *displacements*
- Step 3: **TCN Classification**
  - 1D TCN as *sequence to sequence* classifier
  - *Grid-search* for best model specification: stack of 7 residual blocks with 128 units
  - Trained with *weighted* cross-entropy loss to account of sparsity of labels
  - *Adam* optimizer with default PyTorch parameters
- Step 4: **Post-process**
  - Off-the-shelf *HMM post-process* [1] to obtain the final beat positions

4. Improvement

- Baseline TCN is trained with **weighted cross-entropy** loss $\mathcal{L}_{ce}$
- Propose a custom loss term $\mathcal{L}_p$ that **improves performance**
- *Idea*: exploit the **periodicity of output** based on ground truth tempo
  - Beat probabilities at interval apart should be considered similar
  - $\mathcal{L}_p$ used only on the training set
  - $\mathcal{L}_p$ is mixed with a parameter α estimated by grid-search and summed to $\mathcal{L}_{ce}$

$$\mathcal{L}_p^{(n)} = \left\| \sum_{\substack{k=0 \\ k'=k+1 \\ k''=k'+1}}^{N_b-3} \hat{\mathrm{y}}_{kT:k'T}^{(n)} - \hat{\mathrm{y}}_{k'T:k''T}^{(n)} \right\|_1$$



$$\mathcal{L} = \mathcal{L}_{ce} + \alpha\mathcal{L}_p$$

$$\mathcal{L}_p \rightarrow \hat{\mathrm{y}}_{0:T} \approx \hat{\mathrm{y}}_{T:2T} \approx \dots$$

$T \coloneqq$ inter-beat interval that corresponds to the ground truth tempo

5. Results

- Test our algorithm on the **AIST Dance Video Database** [2]
  - Use the *subset of videos* recorded by the frontal camera and feature one dancer a time
  - Consider data splits based on "dancer" and "music"
  - Randomly split: 70% training, 20% validation, 10% test
- Dance beat tracking is a **challenging** task
  - Performance results are lower if compared to music beat tracking
  - The performance on the "dance" split is higher than the performance on the "music" split
- The periodicity loss achieves a considerable **improvement** of performance

Loss	CML$_c$	CML$_t$	AML$_c$	AML$_t$	Cem	F
L_{ce}	44.28	46.93	47.27	49.04	52.92	55.02
$L_{ce}+\alpha L_p$	**53.05**	**54.30**	**55.23**	**57.64**	**59.02**	**61.20**

Table 1. Performance results on the "dancer" data split using the proposed loss with $\alpha = 0.05$

Loss	CML$_c$	CML$_t$	AML$_c$	AML$_t$	Cem	F
L_{ce}	40.14	39.71	44.84	47.53	47.43	53.02
$L_{ce}+\alpha L_p$	**46.50**	**48.33**	**48.27**	**50.87**	**54.27**	**58.25**

Table 2. Performance results on the "music" data split using the proposed loss with $\alpha = 0.1$

6. Contributions

- Propose the **novel task** of dance beat tracking using visual information alone
- Propose the **periodicity loss term**, which is scaled and added to the baseline cross-entropy loss
- Provide a **baseline evaluation** on the AIST Dance Video Database considering data splits based on music and dancer

7. References

[1] F. Krebs, S. Böck and G. Widmer, "An Efficient State Space Model for Joint Tempo and Meter Tracking", ISMIR 2015

[2] S. Tsuchida, S. Fukayama, M. Hamasaki and M. Goto, "AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing", ISMIR 2019