

Defensive Publication Bundle — v1.0.0

Part A — Defensive Publication (Prior Art) v1.0.0

Title (EN)

On-device Temporal Alignment of Two Practice Videos via Pose-Keypoint Sequence Matching with Optional Audio-Assisted Candidate Generation and/or Audio Verification / Micro-Refinement

タイトル (JP)

骨格キー・ポイント・系列照合による2本の練習動画の開始点自動同期（端末内）－音声支援の粗密化・検証ハイブリッドおよび探索高速化（Phase1例示）を含む

Abstract (EN)

This disclosure describes an on-device method to automatically estimate the temporal offset between two practice videos containing the same or highly similar performance segment (e.g., dance/ballet), enabling synchronized side-by-side playback.

The method samples frames at a fixed FPS and extracts pose keypoints per frame using an on-device pose estimator. The disclosure covers both single-person and multi-person pose estimators. For multi-person outputs, a target person can be selected using detection score, average keypoint confidence, spatial consistency, track continuity, and/or region-of-interest constraints.

Keypoints are normalized into a pose feature vector by filtering low-confidence points, selecting a robust center (pelvis-first, then shoulders, then centroid fallback), and applying scale normalization (shoulder width, hip width, or maximum radius fallback). Temporal alignment is obtained by scanning candidate offsets within \pm searchRangeSec and computing an offset score as the average pose similarity across aligned frames. Per-frame similarity may be cosine similarity, and distance can be defined as $1 - \cosSim$. Frame pairs with fewer than a minimum number of commonly valid keypoints (shared-

valid points) are excluded from scoring to improve robustness.

Offset sign convention: Let Δ denote the shift applied to Video B relative to Video A when pairing frames for scoring. For a frame index t in A, the paired index in B is $(t + \Delta)$. Thus, $\Delta > 0$ indicates that B lags A (B is evaluated later than A), and $\Delta < 0$ indicates that B leads A.

The selected offset may be converted into a robust start point using the earliest contiguous above-threshold similarity segment (or equivalent window-based criteria), with staged relaxation retries if no such segment is found.

A confidence score determines acceptance:

```
```text
confidence = 0.5·bestSim + 0.3·margin + 0.2·validRate
```
```

```

Here, margin is the gap between best and runner-up offset scores, and validRate denotes a validity ratio computed over the matched segment/window (e.g., valid-frame ratio and/or valid-keypoint ratio). When distance-minimization is used internally, bestSim and runner-up values for margin can be computed on any monotonic transform consistent with the acceptance logic (e.g., converting distance to similarity), without loss of generality.

Optional hybrids are also disclosed:

- Audio-assisted coarse candidate generation (top-K offsets with minimum separation) followed by pose refinement within local neighborhoods, and/or
- Post-pose audio verification or micro-refinement around the selected offset to improve precision.

Additionally, an optional pose-only phase acceleration is disclosed (e.g., restricting search to early segments or high-motion windows, with fallback to full search).

This document is published to establish prior art for pose-sequence-based temporal alignment, robust start selection via contiguous similarity, confidence-based rejection,

audio-assisted coarse-to-fine refinement and/or audio verification, and mobile-efficient phase acceleration.

**\*\*Keywords:\*\*** pose estimation, skeleton keypoints, temporal alignment, time offset estimation, sequence matching, cosine similarity, contiguous match, confidence scoring, coarse-to-fine, audio-to-pose, audio verification, micro-refinement, on-device, mobile, multi-pose

---

## ## 概要 (JP)

本開示は、同一または高類似の演技区間を含む練習動画2本（例：バレエ／ダンス）を左右に並べて比較する際に、両動画の開始位置（時間オフセット）を端末内で自動推定し、同期再生に用いる方式を示す。

手法は、一定FPSでフレーム列を作成し、各フレームで骨格キーポイント（座標+信頼度）を推定する。單一人物推定に限らず、複数人物推定（Multi-Pose）の出力から対象人物を選別して用いる形態を含む（例：検出スコア、平均キーポイント信頼度、空間的一貫性、追跡の連續性、ROI制約など）。

低信頼キーポイントを除外した上で、中心（骨盤優先→肩→重心フォールバック）とスケール（肩幅／腰幅／最大半径フォールバック）により正規化した骨格ベクトル系列を構築する。次に、 $\pm\text{searchRangeSec}$  の範囲で時間オフセットを走査し、対応フレーム間の骨格ベクトル類似度（例：コサイン類似度）を平均集計したスコアで最良オフセットを選ぶ（距離は  $1-\cos\text{Sim}$  と定義可能）。この際、両フレームで共通して有効なキーポイント数が所定数未満のフレーム対は、スコア計算対象から除外し頑健性を高める。

**\*\*符号規約：\*\***  $\Delta$  を「Video A に対して Video B 側を評価する時間（フレーム）シフト」とし、Aのフレームtに対しBのフレーム(t+ $\Delta$ )を対応づける。 $\Delta>0$  は「BがAより遅れている（BはAより後の時刻として評価される）」、 $\Delta<0$  は「BがAより先行している」を意味する。

最良オフセットは、類似度が閾値以上となる連續区間（連續一致）を用いて頑健な開始点へ変換し、該当区間が見つからない場合は段階的に緩和して再試行する。

採否判定には以下の信頼度を用いる：

```
```text
confidence = 0.5×bestSim + 0.3×margin + 0.2×validRate
```
```

```

ここで margin は最良スコアと次点スコアの差、validRate は一致区間／評価窓における有効率（例：有効フレーム率および／または有効キーポイント率）である。内部で距離最小化を用いる場合、margin は距離→類似度など単調変換後の値で算出してもよい。

加えて、以下のオプション形態も含む：

- (i) 音声類似度で粗い候補オフセットを複数（Top-K、近接ピーク重複回避の最小間隔つき）生成し、その近傍のみをPose系列照合で精密化する粗密ハイブリッド（音声→Pose）
- (ii) Poseで得たオフセットを音声で検証・微調整（micro refinement）する形態
- (iii) 端末高速化のために、開始区間限定探索や動き量の大きい区間検出に基づく探索窓限定（Phase1）を行い、失敗時はフル探索へフォールバックする形態

本開示は、骨格系列に基づく動画時間同期、信頼度による棄却、音声支援の粗密精密化および／または音声検証・微調整、ならびに端末向け探索高速化を先行技術化する目的で公開する。

A.1 Problem Setting / 課題設定

Two practice videos may contain the same or highly similar performance segment but start at different times. The goal is to estimate a temporal offset (start alignment) for synchronized side-by-side playback on mobile devices, robust to noise such as:

- different camera distance/zoom and minor viewpoint changes,
- partial occlusions and low-confidence keypoints,
- timing differences (pauses, warm-up motions, idle segments),
- pose-estimation noise and missing keypoints.

A.2 Inputs and Outputs / 入出力

Inputs / 入力

- Video A, Video B
- Parameters (examples; not limiting):
 - `F`: sampling FPS (frames per second)
 - `searchRangeSec`: maximum offset search range (\pm , seconds)
 - `τ_kp`: keypoint confidence threshold
 - `M_shared`: minimum number of shared-valid keypoints for a frame-pair to be scored (e.g., ≥ 2 ; implementation-defined)
 - `τ_sim`: similarity threshold for contiguous match detection
 - `N`: minimum contiguous frames above threshold
 - `τ_conf`: acceptance threshold for confidence
- Hybrid parameters (optional):
 - `k`: top-K audio candidates
 - `minSepSec`: minimum separation between candidate peaks (seconds)
 - `r_frames` / `r_sec`: local neighborhood radius around candidates (frames / seconds)
 - `ε_frames` / `ε_sec`: micro-refinement radius for audio verification (frames / seconds)
- Optional runner-up separation (optional):
 - `δ_excl_frames` / `δ_excl_sec`: exclusion radius around Δ^* when selecting runner-up

Outputs / 出力

- `startA_sec`, `startB_sec`: aligned start timestamps for synchronized playback
- `confidence`: acceptance score (e.g., normalized to $[0,1]$)
- Optional diagnostics:
 - `bestOffset_frames`
 - `bestOffset_sec`
 - `bestSim`
 - `secondBestSim`

- `margin`
- `validRate`

Units / 単位 (重要)

Offsets and neighborhoods can be represented in either frames or seconds. In this disclosure, **** Δ is expressed in frames by default**** (integer frame shift), and its seconds representation is:

```
```text
Δ_sec = Δ / F
r_sec = r_frames / F
ε_sec = ε_frames / F
```
---
```

Equivalent embodiments may instead express all quantities directly in seconds.

A.3 Method / 手法

A.3.1 Frame Sampling / フレーム抽出

1. Decode each video into frames sampled at fixed FPS `F` and a fixed target size.
2. Frame index `t` corresponds to timestamp `t/F`.

A.3.2 Pose Estimation / 骨格推定 (Single-pose / Multi-pose)

For each frame, estimate pose keypoints with coordinates and confidence `(x_i, y_i, s_i)`.

- ****Single-pose:**** directly use the single-person output.
- ****Multi-pose:**** select a target person from multiple detected persons using one or more of:
 - detection score / instance score,
 - average keypoint confidence,
 - spatial consistency (e.g., torso center continuity, size constraints),

- track continuity over time (simple tracking by nearest center / IoU / embedding),
- region-of-interest constraints,
- rejecting implausible skeleton geometry.

This disclosure is not limited to any specific model or framework; any on-device pose estimator producing keypoints with confidence qualifies as an equivalent embodiment.

A.3.3 Pose Normalization & Vectorization / 正規化ベクトル化

Purpose: reduce variation due to camera translation and scale.

- Keypoint filtering: $K_{\text{valid}}(t) = \{ i \mid s_i(t) \geq \tau_{\text{kp}} \}$
- Center selection (fallback): pelvis/hip center \rightarrow shoulder center \rightarrow centroid of valid keypoints
- Scale selection (fallback): shoulder width \rightarrow hip width \rightarrow max radius from center
- Normalized coordinates (for valid keypoints):
 - $p_i(t) = ((x_i(t) - c_x(t)) / \alpha(t), (y_i(t) - c_y(t)) / \alpha(t))$
- Feature vector: concatenate normalized coordinates in a fixed keypoint order

Missing keypoints may be handled by masking (shared-valid only), zero-imputation with mask, or variable-length aggregation normalized by count.

A.3.4 Offset Scan & Score / オフセット走査とスコア

Offset sign convention: For frame index t in A, pair with $(t + \Delta)$ in B. Thus, $\Delta > 0$ means B lags A**, and $\Delta < 0$ means B leads A**.

Search range (frames):

- $\Delta \in [-\text{searchRangeSec}\cdot F, +\text{searchRangeSec}\cdot F]$

Discrete-range note (implementation-defined): since $\text{searchRangeSec}\cdot F$ may not be an integer, the conversion from seconds to frame bounds may apply rounding such as `round`, `floor`, or `ceil`, depending on implementation.

Shared-valid minimum:

- compute $\text{sharedValidCount}(t, \Delta)$ = number of keypoints valid in both frames
- exclude frame pairs with $\text{sharedValidCount}(t, \Delta) < M_{\text{shared}}$

- `M_shared` is implementation-defined (e.g., ≥ 2) and tunable

Per-frame similarity:

- example: `sim(t, Δ) = cos(vA(t), vB(t+Δ))`
- distance form: `dist(t, Δ) = 1 - sim(t, Δ)`

Offset score:

- `Score(Δ) = average_t sim(t, Δ)` over included pairs (or minimize average distance)

Best and runner-up:

- $\Delta^* = \operatorname{argmax} \operatorname{Score}(\Delta)$
- $\Delta^2 = \operatorname{argmax}_{\{\Delta \neq \Delta^*\}} \operatorname{Score}(\Delta)$
- optional runner-up separation: exclude $|\Delta - \Delta^*| \leq \delta_{\text{excl_frames}}$ to avoid near-duplicate peaks

Optional (non-limiting): enforce a minimum overlap length and discard too-short overlaps.

A.3.5 Robust Start Point via Contiguous Match / 連続一致区間による開始点

- Find earliest `t0` such that `sim(t, Δ*) ≥ τ_sim` holds for at least `N` consecutive frames (valid frame pairs only).
- If not found, staged relaxation retries may lower `τ_sim`, reduce `N`, switch to window-ratio criteria, switch to distance thresholds, etc.

Aligned start timestamps:

- Global search:
 - `startA_sec = t0 / F`
 - `startB_sec = (t0 + Δ*) / F`
- Windowed/local search (seeded neighborhoods, Phase1 restricted windows):
 - `startA_sec = baseA_sec + t0 / F`
 - `startB_sec = baseB_sec + (t0 + Δ*) / F`

A.3.6 Confidence & Acceptance / 信頼度と採否

Definitions:

- `bestSim = Score(Δ*)`

- `secondBestSim = Score(Δ2)`
- `margin = bestSim - secondBestSim`

Validity ratio `validRate` (over matched segment/window):

- `validFrame(t, Δ)` = 1 if the pair is eligible and included (e.g., `sharedValidCount≥M_shared`), else 0
- `validFrameRate` = average of `validFrame(t, Δ*)`
- optional: include validKeypointRate; validRate may denote valid-frame ratio and/or a combination

Confidence (example):

```
```text
confidence = 0.5 × bestSim + 0.3 × margin + 0.2 × validRate
```

```

Accept if `confidence ≥ τ_conf`, else reject.

Distance-minimization note:

- if computed via distance minimization internally, `bestSim`/`secondBestSim` for margin may be computed after any monotonic transform (e.g., distance→similarity)
- values may be clipped/normalized to [0,1] prior to thresholding

A.4 Optional Hybrids / オプション：音声支援の粗候補生成および／または音声検証・微調整

A.4.1 Audio-based coarse candidate generation (optional) / 音声で粗候補生成（任意）

Compute audio similarity using any audio features (energy envelope, onset patterns, spectral features, fingerprint, cross-correlation, etc.). Generate top-K candidate offsets `Δc_1..Δc_k` (in frames or seconds).

Optional robustness:

- enforce minimum separation between candidate peaks `minSepSec` to avoid near-

duplicate candidates,

- keep both positive and negative candidates (respecting Δ sign convention).

A.4.2 Pose refinement around candidates (optional) / Poseで候補近傍を精密化（任意）

For each candidate, refine only within a local neighborhood:

- $\Delta \in [\Delta_{c_i} - r_{\text{frames}}, \Delta_{c_i} + r_{\text{frames}}]$
(or seconds: $\Delta_{\text{sec}} \in [\Delta_{c_{\text{sec}}} - r_{\text{sec}}, \Delta_{c_{\text{sec}}} + r_{\text{sec}}]$)

Compute $\text{Score}(\Delta)$ using pose matching (Section A.3), select best Δ^* across all neighborhoods, then apply contiguous-match and confidence.

A.4.3 Post-pose audio verification / micro refinement (optional) / Pose後の音声検証・微調整（任意）

After a pose-derived offset Δ^* is selected:

- evaluate audio similarity in a small neighborhood $\Delta \in [\Delta^* - \varepsilon_{\text{frames}}, \Delta^* + \varepsilon_{\text{frames}}]$
(or seconds)
 - if a nearby offset yields better audio consistency (or passes an audio check), adjust Δ^* accordingly
 - or reject if audio strongly contradicts the pose-derived alignment

A.5 Optional Pose-only Phase Search Acceleration (Phase1; Android-style example) / Poseのみの探索高速化（Phase1；Android風の例示）

To improve runtime on mobile devices, an optional pose-only acceleration phase can be used:

- **Early-segment restricted search:** search only within early segments (e.g., 0–10s, 10–20s) before expanding.
- **High-motion window selection:** detect windows with high motion magnitude and search only those windows. Examples of motion proxies include pose displacement norms across frames, simple frame-difference energy, optical-flow statistics (if available), and other lightweight motion proxies.
- **Fallback:** if no confident match is found in restricted windows, fall back to the full-

range scan (Section A.3.4).

This phase acceleration is orthogonal to audio-assisted hybrids and can be combined with them.

A.6 Complexity & Mobile Considerations / 計算量と端末上の考慮

- Pose inference dominates runtime; reducing FPS and input size is effective.
- Candidate neighborhood refinement reduces offset evaluations.
- Shared-valid filtering, normalization, and validRate/margin improve robustness.

A.7 Variations / Equivalent Embodiments / 変形例（同等実施形態）

This disclosure explicitly includes (non-exhaustive):

1. similarity metrics (cosine, L2, weighted, joint-angles, masked),
2. search strategies (brute-force, multi-resolution, coarse-to-fine, beam, dynamic alignment),
3. robust start detection (run-length, window ratio, hysteresis, relaxation),
4. confidence scoring (linear example, rule-based, uncertainty),
5. normalization variants and smoothing,
6. multi-person target selection and tracking variants,
7. hybrid designs (audio→pose, pose→audio, manual seed→pose).

A.8 Failure Handling and Fallbacks / 失敗時のフォールバック

If pose alignment fails or `confidence < τ_conf`:

- fall back to audio-only alignment (if available), and/or
- keep the pose offset but require audio verification before acceptance, and/or
- reject automatic alignment and allow manual adjustment.

A.9 Intended Use as Prior Art / 先行技術化の意図

This document is published to establish prior art for:

- pose-keypoint sequence temporal alignment with bounded offset scan,
- shared-valid minimum exclusion for low-quality frame pairs,
- robust start-point selection via contiguous similarity (with relaxation),
- confidence-based rejection using bestSim/margin/validRate (example coefficients included),
- audio-assisted top-K candidate generation and local pose refinement (audio→pose),
- post-pose audio verification/micro refinement (pose→audio),
- mobile-efficient phase acceleration (Phase1) with fallback.

A.10 Publication Metadata / 公開メタ情報

- ****Authors / 著者**** : SparklyForce (または希望表記)
- ****Version**** : v1.0.0
- ****Publication date (UTC) / 公開日 (UTC) **** : YYYY-MM-DD
- ****First public timestamp (UTC, ISO 8601)**** : YYYY-MM-DDThh:mm:ssZ
- ****Repository tag (optional)**** : v1.0.0
- ****Commit hash (Git SHA)**** : <to be filled>
- ****Permanent URL**** : GitHub Release URL (後で追記)
- ****Immutable archive URL**** : Zenodo record URL / Internet Archive URL (後で追記)
- ****DOI**** : Zenodo Version DOI (発行後に追記)
- ****Hash (optional)**** : PDF SHA-256 (推奨)
