IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

A Physiotherapy Video Matching Method Supporting Arbitrary Camera Placement via Angle-of-Limb-based Posture Structures

Jiunn-Wu Lin[1], Yao-Sheng Chou[2], Yun-Pin Huang[3], Min-Hsiung Hung[4], *Senior Member, IEEE* Ming-Hung Kao[2], Jia Ji[2], Lin-Yi Jiang[2†], Pi-Wei Chen[2], Chao-Chun Chen[2†]

***Abstract*—The "Hospital at Home" initiative transforms medical service automation through modern technologies. This paper revisits remote physiotherapy, allowing convalescents to record exercises using mobile devices from arbitrary angles. To address this, we propose a physiotherapy video matching method that accurately aligns movements from unconstrained viewpoints. The task is formulated as an optimization problem and solved using a modular pipeline. We introduce the Angle-of-Limb-based Posture Structure (ALPS) and the Camera-Angle-Free (CAFE) transformation to counter camera-angle differences. We also develop the Three-phase ALPS Matching Algorithm (TALMA) for matching movements between mentor and convalescent videos. Real-world experiments show our method outperforms existing solutions in both precision and practicality, with a time deviation of less than 0.07 seconds from expert annotations. The prototype and datasets are publicly available at: https://github.com/NCKU-CIoTlab/TALMA-on-ALPS/.**

Keywords: rehabilitation modeling, applied deep learning, telephysiotherapy automation, semantic matching, hospital at home.

I. Introduction

The "Hospital at Home" model has gained increasing traction as a healthcare policy, particularly in countries facing declining birthrates and aging populations[1], [2]. This model aims to deliver hospital-level care in residential settings, leveraging advancements in information technology and automation. In response, research communities in robotic and automation have been actively developing core technologies to enable such remote healthcare applications.Among various medical departments, physiotherapy stands out as a promising pioneer for hospital-at-home implementation, primarily due to its suitability for computer vision-based analysis [3], [4], [5], [6].

Successful development of hospital-at-home applications requires seamless integration of modern IT and AI technologies. This paper explores a remote physiotherapy application that automatically matches movements between video clips—one recorded by a physiotherapy mentor and the other by a convalescent patient at home. Physiotherapy video matching (PVM) aims to identify frames in the patient's clip that best correspond to the selected anchor frames in the mentor's clip. In this study, an anchor frame refers to a manually chosen key frame in the mentor's video representing a critical pose in the physiotherapy motion. By automating this process, healthcare professionals can efficiently review convalescent videos without performing time-consuming, frame-by-frame inspections.

However, PVM introduces two significant technical challenges. The first involves interpreting physiotherapy movements recorded from varied camera angles. Since home-based patients may place their mobile devices differently than mentors, the resulting viewpoint discrepancies complicate visual matching. Thus, subtle differences between similar and dissimilar frame pairs are difficult to distinguish through 2D images alone. This challenge becomes even pronounced in long videos captured from highly divergent angles.

The second challenge concerns the computational difficulty of selecting an appropriate frame subset that semantically matches the mentor's movements. This task must consider physiotherapy-specific constraints, such as avoiding many-to-one matches and accounting for unstable similarity metrics caused by view variation. These complexities make PVM a nontrivial and demanding problem.

Although prior work on physiotherapy movement analysis has demonstrated promising results in posture and joint-angle evaluations using predefined angles, such as Q-angle assessments [7] and single-plane analyses [3], these methods assume fixed viewpoints and thus struggle with view-invariant recognition. Several approaches have attempted to improve consistency using software-assisted assessments [4] and motion analysis systems [8], but all remain constrained by static camera perspectives. Additional efforts, including the movement variability framework [9] and postural scoring systems [10], provide valuable insights under controlled conditions but do not support customizable or view-invariant evaluations.

The sports domain has significantly progressed in fine-grained action parsing and temporal alignment. For example, the FineGym [11] supports hierarchical temporal analysis of complex gymnastics routines, while the KIMORE [12] enables clinical evaluation of therapeutic movements. The LOGO [13] targets group activity understanding with interpersonal modeling. Despite their technical advancements, these works rely heavily on fixed or controlled viewpoints, offering limited support for cross-view matching or adaptable assessment.

Manuscript received: January 28, 2025; Revised April 13, 2025; Accepted June 19, 2025. This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments. Authors thank K. Wang and B.-H. Chiu with IMIS, NCKU for developing the system prototype. This work was supported by Taiwan's National Science and Technology Council (NSTC; Grants 113-2221-E-006-125, 113-2221-E-034-002-MY2). †Corresponding authors: Lin-Yi Jiang and Chao-Chun Chen.

[1]J.-W. Lin is with Information Management Office, Kaohsiung Veterans General Hospital, Taiwan. (e-mail: jiunnwu@vghks.gov.tw)

[2]Y.-S. Chou, M.-H. Kao, J. Ji, L.-Y. Jiang†, P.-W. Chen, and C.-C. Chen† are with Institute of Manufacturing Information and Systems, Department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Taiwan; (e-mail: p98101500@gs.ncku.edu.tw, chencc@imis.ncku.edu.tw).

[3]Y.-P. Huang is with Playsure Inc., Taiwan. y0918888678@gmail.com

[4]M.-H. Hung is with the Department of Computer Science and Information Engineering, Chinese Culture University, Taiwan. (e-mail: hmx4@ulive.pccu.edu.tw). Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Across rehabilitation and sports analytics, existing methods fail to address two critical needs: robustness to camera viewpoint variation and support for dynamic, customizable evaluation criteria. This gap motivates our work to design a camera-angle-invariant and semantically flexible framework for physiotherapy video matching.

To address these challenges, we propose a novel PVM method based on three key innovations:

(1) Optimization-based Modular Pipeline: We reformulate PVM as an optimization problem with semantic constraints and design a modular pipeline comprising pose acquisition, structural transformation, and semantic alignment stages. To efficiently process input data, we adopt off-the-shelf models, Alphapose [5] for 2D keypoint detection and DST [6] for 3D reconstruction, as components within our custom-defined modules, PHP-Net and 3DPHP-Net.

(2) Angle-of-Limb-Based Semantic Representation: We introduce the Angle-of-Limb-based Posture Structure (ALPS), which captures human posture through inter-limb angles, providing strong invariance to camera viewpoints. To derive ALPS representation from arbitrary perspectives, we propose the Camera-Angle-Free Encoding (CAFE) transformation, which converts 2D/3D keypoints into a unified, view-agnostic format.

(3) Three-Phase ALPS Matching Algorithm (TALMA): We develop TALMA to perform progressive, structure-aware matching of physiotherapy movements. This algorithm addresses the NP-complete nature of the matching problem using heuristic techniques that balance computational efficiency with alignment accuracy, making it suitable for practical applications.

We evaluate our method in real-world tele-physiotherapy, where a nurse performs 11 standard exercises. The system matches patient videos with an average temporal alignment error of less than 0.07 seconds compared to expert annotations.

II. Problem Formulation and Sketch of Solution

To facilitate understanding, Fig. 1 illustrates the symbolic convention used for human posture vectors in this paper. Each posture vector is written as h (e.g., ${}^{\theta}_{r}\mathbf{h}^{i}_{j}$)[1], where θ denotes the camera angle[2], r represents the owner role of $\mathbf{h}$ (either a mentor m or a convalescent c), i is the frame index, and j is the vector element index.

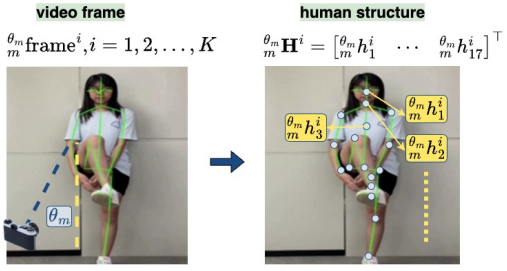


Fig. 1. Mapping between symbolic definitions and a physiotherapy video frame.

[1]Superscripts and subscripts are used consistently across symbols serving similar purposes.

[2]In this paper, the term "camera angle" refers to the observed orientation of a subject's pose as captured from a specific camera placement. It is primarily influenced by the camera's position and direction relative to the subject.

A video clip comprises a sequence of ordered frames. For the mentor, we denote the ordered frame set as ${}^{\theta_m}_{m}\mathcal{V} = \{{}^{\theta_m}_{m}f^{i} | i = 1, \ldots, M\}$, where ${}^{\theta_m}_{m}f^{i}$ represents the i-th frame captured at camera angle θ_m. Similarly, for the convalescent, the video is represented as ${}^{\theta_c}_{c}\mathcal{V} = \{{}^{\theta_c}_{c}f^{i} | i = 1, \ldots, N\}$, where ${}^{\theta_c}_{c}f^{i}$ denotes the i-th frame recorded at angle θ_c.

The mentor selects $K \in \mathbb{N}$ anchor frames from ${}^{\theta_m}_{m}\mathcal{V}$ to serve as reference poses for evaluation. We denote the indices of these frames by $\alpha_i \in \mathbb{N}$, where $i = 1, \ldots, K$, and $1 \leq \alpha_i \leq N$, with N representing the total number of frames in the mentor's video. Each anchor frame encapsulates a target posture that the mentor expects the convalescent to imitate.

To evaluate movement similarity, the system identifies the most relevant matching frames from the convalescent's video ${}^{\theta_c}_{c}\mathcal{V}$ for each anchor frame in ${}^{\theta_m}_{c}\mathcal{V}$.It then computes a score based on transformed posture representations, capturing the body configuration's structural and directional similarity. A higher score indicates a closer match between the convalescent's pose and the mentor's reference. This scoring mechanism enables interpretable and quantifiable assessments of motion imitation quality across all anchor frames.

At the core of our approach is the Physiotherapy Video Matching (PVM) method, which selects a subset of frames from ${}^{\theta_c}_{c}\mathcal{V}$ that optimally align with the mentor's designated anchor frames $\{{}^{\theta_m}_{m}f^{\alpha}_{i} | i = 1, \ldots, K\}$.By automating this process, our PVM method allows physiotherapists to efficiently assess the rehabilitation progress of many convalescents without manually reviewing videos frame by frame.

This capability significantly reduces workload, improves consistency, and supports large-scale or tele-physiotherapy programs. We define the PVM problem as follows.

PVM Problem. *Let $\alpha_i, i = 1, \ldots, K$, be the index of the anchor frames. Find K frame index β_j from ${}^{\theta_c}_{c}\mathcal{V}$, denoted by $\hat{Z}^* = \{\beta_j | j = 1, \ldots, K$ and $1 \leq \beta_j \leq N\}$, such that the selected K frames are most similar to the corresponding anchor frames in terms of movement measures under the following requirements in remote physiotherapy scenarios:*

1) The physiotherapy movement comparison between frames shall avoid interference from visual appearance characteristics, such as weight, height, clothes, etc.;

2) The physiotherapy video clips for the mentor and convalescents can be shot from arbitrary camera angles.

For computationally studying the issue, the problem can be represented in an optimization form:

$$\textit{Maximize} \quad \underbrace{sim\left(\left\{{}^{\theta_m}_{m}f^{\alpha_i}\right\}_{i=1}^{K}, \left\{{}^{\theta_c}_{c}f^{\beta_j}\right\}_{j=1}^{K}\right)}_{\textbf{Need to satisfy requirements (1) and (2).}} \tag{1}$$

The solution $\hat{Z}^*$ to Eq.(1) is thus expressed in the equation:

$$\hat{Z}^* = \arg\max_{\{\beta_i\}} \sum_{i=1}^{K} sim({}^{\theta_m}_{m}f^{\alpha_i}, {}^{\theta_c}_{c}f^{\beta_i}) \tag{2}$$

Directly solving the PVM problem by comparing the similarity between video clips ${}^{\theta_m}_{m}\mathcal{V}$ and ${}^{\theta_c}_{c}\mathcal{V}$ presents a significant challenge. This is because the PVM problem is essentially an NP-complete problem [14], further complicated by domain-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

specific constraints related to human posture, camera angles, and movement semantics.

To address this complexity, we draw inspiration from classic mathematical techniques, such as the Fourier Transform, which reduces the computational cost of time-domain signal processing by operating in the frequency domain. Similarly, we tackle the PVM problem by *transforming the video data into a camera-angle-free domain, where matching becomes more tractable.*.

We assume that, for each frame i, an ideal data model ${}_r\mathcal{A}^i$ exists for role r ($r \in \{m, c\}$) to represent human postures in a unified, appearance-independent, and camera-angle-free manner. With this transformation, we reformulate the PVM problem in the following form:

$$\hat{Z}^* = \arg\max_{\{\beta_i\}} \sum_{i=1}^{K} \underbrace{\text{sim}({}_m\mathcal{A}^{\alpha_i}, {}_c\mathcal{A}^{\beta_i})}_{\text{(1) appearance-independent structure, (2) no camera angles.}} \tag{3}$$

We incorporate the two requirements specified in the PVM problem into the design of the data model ${}_r\mathcal{A}^i$. To bridge Eq.(2) and Eq.(3), we introduce three abstract functions as solution-representation tools to systematically construct ${}_r\mathcal{A}^i$. The first abstract function, $\mathfrak{F}_{\text{PHP}}$, extracts a two-dimensional (2D) keypoint-based structure known as the Purified Human Posture (PHP) model. This model, denoted by ${}^{\theta_r}_{r}\mathbf{H}^i$, represents the purified posture of the i-th frame for role r, captured from camera angle θ_r, and is defined as follows:

$${}^{\theta_r}_{r}\mathbf{H}^i \triangleq \mathfrak{F}_{\text{PHP}}({}^{\theta_r}_{r}\text{f}^i) \tag{4}$$

The second abstract function, $\mathfrak{F}_{\text{3DPHP}}$, performs an up-projection of the 2D PHP model ${}^{\theta_r}_{r}\mathbf{H}^i$ into a 3D PHP model ${}^{\theta_r}_{r}\mathbf{Q}^i$, defined as follows:

$${}^{\theta_r}_{r}\mathbf{Q}^i \triangleq \mathfrak{F}_{\text{3DPHP}}({}^{\theta_r}_{r}\mathbf{H}^i) \tag{5}$$

The resulting 3D PHP structure ${}^{\theta_r}_{r}\mathbf{Q}^i$ enables the model to effectively mitigate the impact of camera-angle variations on human posture representation. The third abstract function $\mathfrak{F}_{\text{CAFE}}$, transforms the 3D PHP structure ${}^{\theta_r}_{r}\mathbf{Q}^i$ into a camera-angle-free representation ${}_r\mathbf{A}^i$ which removes all dependencies on camera angle parameters. This function is defined as:

$${}_r\mathbf{A}^i \triangleq \mathfrak{F}_{\text{CAFE}}({}^{\theta_r}_{r}\mathbf{Q}^i) \tag{6}$$

We implement the ideal data model ${}_r\mathbf{A}^i$ using the output of $\mathfrak{F}_{\text{CAFE}}$ thereby grounding the theoretical formulation in a practical representation.

By the composition of the above three abstract functions, $\mathfrak{F}_{\text{PHP}}$, $\mathfrak{F}_{\text{3DPHP}}$ and $\mathfrak{F}_{\text{CAFE}}$, we derive the final PVM solution form introduced in Eq.(2) as follows:

$$\begin{aligned}
\hat{Z}^* &= \arg\max_{\{\beta_i\}} \sum_{i=1}^{K} \text{sim}({}^{\theta_m}_{m}\text{f}^{\alpha_i}, {}^{\theta_c}_{c}\text{f}^{\beta_i}) \\
&= \arg\max_{\{\beta_i\}} \sum_{i=1}^{K} \text{sim}(\mathfrak{F}_{\text{CAFE}} \circ \mathfrak{F}_{\text{3DPHP}} \circ \mathfrak{F}_{\text{PHP}}({}^{\theta_m}_{m}\text{f}^{\alpha_i}), \\
&\qquad\qquad \mathfrak{F}_{\text{CAFE}} \circ \mathfrak{F}_{\text{3DPHP}} \circ \mathfrak{F}_{\text{PHP}}({}^{\theta_c}_{c}\text{f}^{\beta_i})) \qquad (7) \\
&= \arg\max_{\{\beta_i\}} \sum_{i=1}^{K} \text{sim}({}_m\mathbf{A}^{\alpha_i}, {}_c\mathbf{A}^{\beta_i}) \ \text{(connecting Eq. (3) to Eq. (2))} \\
&= \underbrace{\arg\min_{\{\beta_i\}} \sum_{i=1}^{K} \left(1 - \cos({}_m\mathbf{A}^{\alpha_i}, {}_c\mathbf{A}^{\beta_i})\right)}_{\text{A combinatorial optimization problem.}} \qquad (8)
\end{aligned}$$

Eq. (8) presents the core of our solution approach: *the PVM solution $\hat{Z}^*$ can be obtained through a combinatorial optimization process, provided that the three abstract functions (Eqs.(4-6)) are instantiated as indicated in Eq. (7).*

III. Proposed Pipeline for Solving PVM

Based on Eqs. (7) and (8), Fig. 2 illustrates the overall pipeline for solving the PVM problem. The pipeline comprises four modules: the first three, i.e., PHP Net, 3DPHP Net, and CAFE Transformation, implement our proposed abstract functions, namely $\mathfrak{F}_{\text{PHP}}$, $\mathfrak{F}_{\text{3DPHP}}$, and $\mathfrak{F}_{\text{CAFE}}$ as defined in Eq. (7). The final module, TALMA, addresses the optimization problem formulated in Eq.(8). This pipeline operates by executing these four modules sequentially, following the function execution order specified in Eq.(7).

We divide the pipeline into three conceptual stages to clarify its operation logic: extraction, transformation, and matching.

- In the extraction stage, PHP Net applies Eq. (4) to extract human keypoints ${}^{\theta_r}_{r}\mathbf{H}^i$ from video clips, thereby capturing posture semantics.
- The transformation stage involves the 3DPHP Net and the camera-angle-free (CAFE) transformation. These modules generate 3D human keypoint ${}^{\theta_r}_{r}\mathbf{Q}^i$ according to Eq. (5) and compute the angle-of-limb-based posture structure (ALPS) ${}_r\mathbf{A}^i$ (to be discussed later) for Eq. (6), enabling the representation of physiotherapy movements under arbitrary camera angles.
- The matching stage incorporates the three-phase ALPS matching algorithm (TALMA), which computes the optimal matching result $\hat{Z}^*$ for the physiotherapy video matching (PVM) problem defined in Eq.(8).

We will provide full technical details of each module in Section IV.

To satisfy requirement (1) of the PVM problem, ensuring that video matching between a mentor and a convalescent remains unaffected by visual appearance, we rely exclusively on pose-relevant features. Fig. 3 presents the human structure definitions employed in our method. The mathematical definitions of the structures shown in Fig. 2 are elaborated as follows. Specifically, Fig. 3(a) illustrates the two-dimensional PHP model ${}^{\theta_r}_{r}\mathbf{H}^i$, constructed from keypoint-based representations and formally defined as follows:

$$\begin{aligned}
&{}^{\theta_r}_{r}\mathbf{H}^i \triangleq \begin{bmatrix} {}^{\theta_r}_{r}\mathbf{h}^i_1 & \cdots & {}^{\theta_r}_{r}\mathbf{h}^i_{17} \end{bmatrix}^\top \in \mathbb{R}^{17\times 2}, \\
&\text{where } {}^{\theta_r}_{r}\mathbf{h}^i_j = \begin{bmatrix} x_j & y_j \end{bmatrix} \in \mathbb{R}^{1\times 2},\ j = 1, \ldots, 17.
\end{aligned} \tag{9}$$

$[x_j\ y_j]$ is the 2D coordinate of keypoint ${}^{\theta_r}_{r}\mathbf{h}^i_j$. The 3D PHP model ${}^{\theta_r}_{r}\mathbf{Q}^j$ follows the same structure definition in Fig. 3(a), except keypoints are raised to 3D vectors, defined as:

$$\begin{aligned}
&{}^{\theta_r}_{r}\mathbf{Q}^i \triangleq \begin{bmatrix} {}^{\theta_r}_{r}\mathbf{q}^i_1 & \cdots & {}^{\theta_r}_{r}\mathbf{q}^i_{17} \end{bmatrix}^\top \in \mathbb{R}^{17\times 3}, \\
&\text{where } {}^{\theta_r}_{r}\mathbf{q}^i_j = \begin{bmatrix} x_j & y_j & z_j \end{bmatrix} \in \mathbb{R}^{1\times 3}\ j = 1, \ldots, 17.
\end{aligned} \tag{10}$$

It is important to note that under varying camera angles, a human posture represented by a keypoint-based model (e.g., ${}^{\theta_r}_{r}\mathbf{H}^i$ and ${}^{\theta_r}_{r}\mathbf{Q}^i$) may exhibit significant value discrepancies. These variations complicate the computation of posture similarity between video frames. To address this issue, we design the angle-of-limb-based posture structure (ALPS), denoted by ${}_r\mathbf{A}^i$, to represent the posture of frame i for role r.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

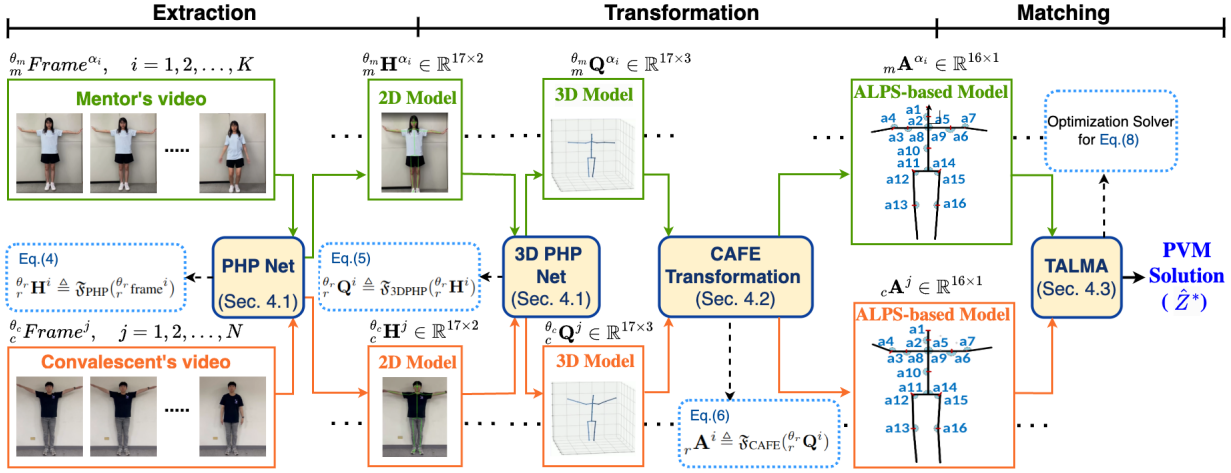


Fig. 2. The proposed physiotherapy video matching pipeline consisting of Extraction, Transformation, and Matching stages.

The ALPS model aims to mitigate the impact of camera angle differences by using posture features that are less sensitive to viewpoint changes, while preserving the most essential semantics of human movements. This design supports both requirements outlined in the PVM problem.

Fig. 3(c) illustrates the ALPS model ${}_r\mathbf{A}^i$, defined as:

$$ {}_r\mathbf{A}^i \triangleq \begin{bmatrix} {}_ra_1^i & \cdots & {}_ra_{16}^i \end{bmatrix}^\top \in \mathbb{R}^{16\times 1} \tag{11}$$

where ${}_ra_j^i$ denotes the angle formed between two adjacent limbs. The detailed formulation and computation of ${}_ra_j^i$ will be provided in Section IV-B.

IV. Key Component Designs

A. Deepnet-empowered Function Implementation: PHP-Net and 3DPHP-Net

To solve the PVM problem efficiently, we integrate Alphapose [5] and DST [6] as pretrained modules within PHP-Net and 3DPHP-Net, respectively—AlphaPose for 2D keypoint extraction and DST for 3D posture reconstruction.

The extracted keypoints are then reformatted to align with the posture structure defined in Fig. 3(a), serving as standardized inputs to the ALPS and CAFE modules. This design streamlines preprocessing, ensures consistency and semantic alignment across the pipeline, and maintains flexibility for future upgrades. Notably, pose estimation backbones such as Alphapose and DST can be replaced or updated without impacting the functionality of the downstream modules.

B. Angle-of-Limb-based Posture Structure (ALPS) and CAFE Transformation

Because keypoint-based features are often sensitive to visual appearance and camera veiwpoint, we instead describe human movements using angle-of-limb vectors. The approach is motivated by the ovservation that limb angles remain relatively invariant across different camera angles, making them more robust for posture comparison.

To construct this representation, we transform the 3D keypoint model ${}^{\theta_r}_{r}\mathbf{Q}^i \in \mathbb{R}^{17\times 3}$ into a limb-based model ${}^{\theta_r}_{r}\mathbf{L}^i \in \mathbb{R}^{16\times 3}$. This transformation is performed by computing each of the 16 limb vectors, ${}^{\theta_r}_{r}\mathbf{l}^i_j$, as the vector difference between a target (end) keypoint vector ${}^{\theta_r}_{r}\mathbf{q}^i_{t_j}$ and a corresponding start keypoint ${}^{\theta_r}_{r}\mathbf{q}^i_{s_j}$, where s_j and t_j are start and end indices of limb j, respectively, based on the keypoint connections defined in Fig. 3(b). The resulting limb-based model is formally defined as:

$$\begin{aligned} &{}^{\theta_r}_{r}\mathbf{L}^i \triangleq \begin{bmatrix} {}^{\theta_r}_{r}\mathbf{l}^i_1 & \cdots & {}^{\theta_r}_{r}\mathbf{l}^i_{16} \end{bmatrix}^\top \in \mathbb{R}^{16\times 3}, \text{ where} \\ &{}^{\theta_r}_{r}\mathbf{l}^i_j = {}^{\theta_r}_{r}\mathbf{q}^i_t - {}^{\theta_r}_{r}\mathbf{q}^i_s = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \in \mathbb{R}^{1\times 3},\ j = 1,\ldots,16. \end{aligned} \tag{12}$$

For example, Limb 1 (Head) connects keypoint 11 (the start point) to keypoint 10 (the end point) and is represented by the vector ${}^{\theta_r}_{r}\mathbf{l}^i_1 = {}^{\theta_r}_{r}\mathbf{q}^i_{11} - {}^{\theta_r}_{r}\mathbf{q}^i_{10}$.

The ALPS vector, denoted as ${}_r\mathbf{A}^i \in \mathbb{R}^{16\times 1}$, is derived from the limb-based representation ${}^{\theta_r}_{r}\mathbf{L}^i$. To compute it, we first normalize each limb vector ${}^{\theta_r}_{r}\mathbf{l}^i_k \in {}^{\theta_r}_{r}\mathbf{L}^i$ to produce a set of unit limb vectors, forming the matrix $\mathbf{U} \in \mathbb{R}^{16\times 3}$. The k-th row, $r_k(\mathbf{U})$, denoted as $r_k(\mathbf{U})$, is defined as:

$$r_k(\mathbf{U}) = \frac{{}^{\theta_r}_{r}\mathbf{l}^i_k}{\|{}^{\theta_r}_{r}\mathbf{l}^i_k\|_2}, \quad \text{for } k = 1,\ldots,16. \tag{13}$$

Next, each component ${}_ra_j^i$ of the ALPS vector ${}_r\mathbf{A}^i$ is computed as the cosine similarity between two specific limb vectors, namely the s_j-th and e_j-th limbs, according to predefined limb parings illustrated in Fig. 3(c). Mathematically, this is expressed as:

$${}_ra_j^i = r_{s_j}(\mathbf{U}) \cdot r_{e_j}(\mathbf{U})^\top = \cos\left({}^{\theta_r}_{r}\mathbf{l}^i_{s_j}, {}^{\theta_r}_{r}\mathbf{l}^i_{e_j}\right), \quad \text{for } j = 1,\ldots,16. \tag{14}$$

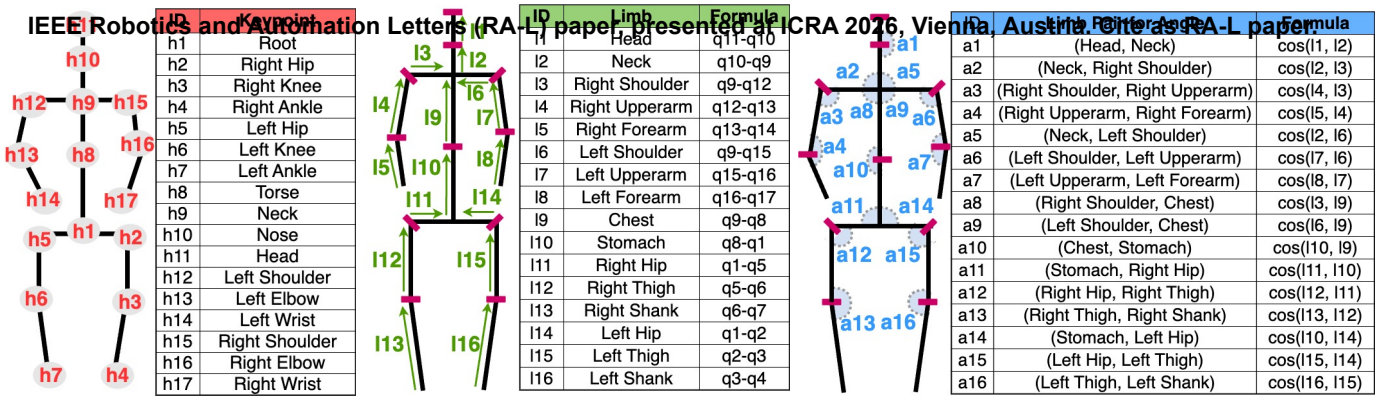
The full ALPS vector ${}_r\mathbf{A}^i$ is assembled as:

$${}_r\mathbf{A}^i = \begin{bmatrix} {}_ra_1^i & {}_ra_2^i & \cdots & {}_ra_{16}^i \end{bmatrix}^\top, \tag{15}$$

Given that Eq. (12) defines the transformation from the 3D keypoints model ${}^{\theta_r}_{r}\mathbf{Q}^i$ to the limb-based model ${}^{\theta_r}_{r}\mathbf{L}^i$, the complete conversion from ${}^{\theta_r}_{r}\mathbf{Q}^i$ (e.g., computed by a deep networks as ${}^{\theta_r}_{r}\mathbf{Q}^i = \text{3DPHP} \circ \text{PHP}({}^{\theta_r}_{r}\text{f}^i)$) to the ALPS representation ${}_r\mathbf{A}^i$ is fully realized through Eqs. (12), (14) and (15). Each element in ${}_r\mathbf{A}^i$ quantitatively encodes the relative angular orientation between two connected limbs. This angle-based representation substantially mitigates sensitivity to absolute camera viewpoints, offering greater robustness.

C. Three-Phase ALPS Matching Algorithm for PVM Solver

To mitigate the effect of occluded or distorted limb angles caused by the varying camera viewpoints, we introduce three variants of the ALPS representation in the matching algorithm:



ID	Keypoint
h1	Root
h2	Right Hip
h3	Right Knee
h4	Right Ankle
h5	Left Hip
h6	Left Knee
h7	Left Ankle
h8	Torse
h9	Neck
h10	Nose
h11	Head
h12	Left Shoulder
h13	Left Elbow
h14	Left Wrist
h15	Right Shoulder
h16	Right Elbow
h17	Right Wrist

(a) Keypoint-Based

ID	Limb	Formula
l1	Head	q11-q10
l2	Neck	q10-q9
l3	Right Shoulder	q9-q12
l4	Right Upperarm	q12-q13
l5	Right Forearm	q13-q14
l6	Left Shoulder	q9-q15
l7	Left Upperarm	q15-q16
l8	Left Forearm	q16-q17
l9	Chest	q9-q8
l10	Stomach	q8-q1
l11	Right Hip	q1-q5
l12	Right Thigh	q5-q6
l13	Right Shank	q6-q7
l14	Left Hip	q1-q2
l15	Left Thigh	q2-q3
l16	Left Shank	q3-q4

(b) Limb-Based

ID	Limb Pair for Angle	Formula
a1	(Head, Neck)	cos(l1, l2)
a2	(Neck, Right Shoulder)	cos(l2, l3)
a3	(Right Shoulder, Right Upperarm)	cos(l4, l3)
a4	(Right Upperarm, Right Forearm)	cos(l5, l4)
a5	(Neck, Left Shoulder)	cos(l2, l6)
a6	(Left Shoulder, Left Upperarm)	cos(l7, l6)
a7	(Left Upperarm, Left Forearm)	cos(l8, l7)
a8	(Right Shoulder, Chest)	cos(l3, l9)
a9	(Left Shoulder, Chest)	cos(l6, l9)
a10	(Chest, Stomach)	cos(l10, l9)
a11	(Stomach, Right Hip)	cos(l11, l10)
a12	(Right Hip, Right Thigh)	cos(l12, l11)
a13	(Right Thigh, Right Shank)	cos(l13, l12)
a14	(Stomach, Left Hip)	cos(l10, l14)
a15	(Left Hip, Left Thigh)	cos(l15, l14)
a16	(Left Thigh, Left Shank)	cos(l16, l15)

(c) Angle-Based

Fig. 3. Structural definitions of human pose representations used in the physiotherapy video matching framework. (a) 2D and 3D keypoint-based pose structures $\mathbf{H}$ and $\mathbf{Q}$; (b) limb structures $\mathbf{L}$, defined by joint pairs individual body segments; (c) limb angle-based features and cosine similarity representations used to construct the alignment vector $\mathbf{A}(ALPS)$.

full-body, left-body and right-body ALPS. The full-body ALPS, denoted by ${}^{F}_{r}\mathbf{A}^{i}$, for frame i is identical to ${}_{r}\mathbf{A}^{i}$. The left-body ALPS, denoted by ${}^{L}_{r}\mathbf{A}^{i}$, is constructed by setting the right-side elements of ${}_{r}\mathbf{A}^{i}$ to zero, specifically: $a_2 = a_3 = a_4 = a_8 = a_{11} = a_{12} = a_{13} = 0$.Conversely, the right-body ALPS, denoted as ${}^{R}_{r}\mathbf{A}^{i}$, sets the left-side elements to zero: $a_5 = a_6 = a_7 = a_9 = a_{14} = a_{15} = a_{16} = 0$.

Our matching process is designed to be frame-wise and alignment-free, meaning each mentor anchor frame is matched individually to the most similar frame in the convalescent clip without requiring sequence synchronization or a shared temporal reference point.

We propose a Three-Phase ALPS Matching Algorithm (TALMA) to solve the PVM problem by aligning ALPS sequences from mentor and convalescent video clips based on specified anchor frames, $\alpha_1, \ldots, \alpha_K$.Inspired by conventional machining procedures [15], which include roughing, refining, and finishing stages, TALMA sequentially performs rough matching, fine matching, and integration to refine the matching result incrementally across multiple ALPS views. Algorithm 1 summarizes the TALMA procedure, and the following provides a conceptual explanation of each phase.

- Phase 1: Rough Matching. In the first phase, we apply Dynamic Time Warping (DTW) [16] to the three ALPS variants (full-body, left-body, right-body) to obtain preliminary matches for each anchor frame α_i, where $i = 1, \ldots, K$. The output of this phase, denoted by ${}^{[F|L|R]}\mathbf{P1R}$[3], may contain many-to-one matchings, meaning multiple convalescent frames can be assigned to a single anchor frame. This rough approximation is an initial candidate set, though the results may not yet meet physiotherapy alignment expectations.
- Phase 2: Fine Matching. The second phase refines the rough matches into one-to-one mappings using two steps:
  Step 2.1: Filter out redundant matches by selecting the frame with the highest full-body ALPS similarity.
  Step 2.2: For matches with insufficient similarity, search for better alternatives within a bounded index range defined neighboring qualified matches. The refined results are recorded as ${}^{[F|L|R]}\mathbf{P2R}$, representing higher-quality, uniquely matched frame pairs.
- Phase 3: Integration. The final phase aggregates the refined matches from all three ALPS models by selecting, for each anchor frame α_i, the matched frames with the highest similarity score across the three sources. The final output is denoted as: $\hat{Z}^* = \{\beta_i | i = 1, \ldots, K\}$, where each β_i represents the optimal matching frame index in the convalescent's clip for anchor frame α_i, and collectively solves the optimization problem defined in Eq.(8).

To asses temporal alignment quality, we define the frame index gap $|\alpha_i - \beta_i|$ as the *time difference* between the mentor anchor frame and its matched convalescent frame. This value serves as a core evaluation metric in our experimental analysis to quantify matching accuracy and consistency.

Algorithm 1 Summarized Three-Phase ALPS Matching Algorithm (TALMA)

/* Initialization */
Construct ALPS similarity matrices ${}^{Type}\mathbf{Z}$ for $Type \in \{F, L, R\}$ between mentor anchor frames α_i and all convalescent frames j, incorporating a temporal decay factor $tmd(i, j)$.
Phase 1: Rough-Matching
foreach $(Type \in \{F, L, R\})$ **do**
  Apply Dynamic Time Warping (DTW) to ${}^{Type}\mathbf{Z}$ to find an initial warping path;
  Backtrack along the path to obtain many-to-one matching sets ${}^{Type}\mathbf{P1R}$;
end foreach
Phase 2: Fine-Matching
Let ${}^{F}\mathbf{P2R}, {}^{L}\mathbf{P2R}, {}^{R}\mathbf{P2R} = \emptyset$; θ_{sim} is a similarity threshold;
foreach $(Type \in \{F, L, R\})$ **do**
  /* Step 2.1: One-to-one Transformation */
  For each anchor α_i, select the single convalescent frame β_i from ${}^{Type}\mathbf{P1R}$ that has the highest full-body similarity $\cos({}^{F}_{m}\mathbf{A}^{\alpha_i}, {}^{F}_{c}\mathbf{A}^{\beta_i})$, forming ${}^{Type}FirstPassMatching$;
  /* Step 2.2: Qualification and Re-matching */
  foreach $(\alpha_i, \beta_i, sim_{\beta_i}) \in {}^{Type}FirstPassMatching$ **do**
    if $sim_{\beta_i} < \theta_{sim}$ **then**
      Add α_i to a list $LowSimAnchor$ for re-matching;
    else
      Re-match anchors in $LowSimAnchor$ by searching for a high-similarity frame β' in a dynamically determined range between the last good match in ${}^{Type}\mathbf{P2R}$ and β_i;
      Append $(\beta', \text{similarity})$ to ${}^{Type}\mathbf{P2R}$;
      Clear $LowSimAnchor$;
      Append (β_i, sim_{β_i}) to ${}^{Type}\mathbf{P2R}$;
    end if
  end foreach
end foreach
Phase 3: Integration
$\hat{Z}^* = \emptyset$. **for** $(i = 1$ *to* $K)$ **do**
  Select the ALPS model type $(TMax \in \{F, L, R\})$ whose corresponding match in ${}^{TMax}\mathbf{P2R}[i]$ has the highest similarity for anchor α_i;
  Let β_i be the frame index from this best match.
  $\hat{Z}^* = \hat{Z}^* \cup \{\beta_i\}$;
end for
return $\hat{Z}^*$;

V. Case Study

A. Environmental Settings and Performance Metrics

We implemented the proposed Physiotherapy Video Matching (PVM) method in Python using the PyTorch framework. To realize the $\mathfrak{F}_{\text{PHP}}$ and $\mathfrak{F}_{\text{3DPHP}}$ functions in our modular pipeline,

[3]The symbol ${}^{[F|L|R]}\mathbf{P1R}$ serves as a shorthand for three seperate data structures: ${}^{F}\mathbf{P1R}$, ${}^{L}\mathbf{P1R}$, and ${}^{R}\mathbf{P1R}$.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

we directly integrated the pretrained models Alphapose [5] and DST [6] without modification. Their outputs are processed through standardized interfaces to ensure modularity and implementation isolation. This design facilitates future extensibility—alternative pose estimation models can be substituted seamlessly without impacting the downstream components. In Algorithm 1, the parameters θ_{sim} and ϕ are empirically set to 0.75 and 15%, respectively. To evaluate performance, we use the Mean Absolute Error (MAE) metric, which quantifies the frame-wise discrepancy between the predicted matches and the ground-truth annotations. To validate the proposed method, we conducted a real-world case study using physiotherapy video data collected in collaboration with a hospital in Tainan, Taiwan. All video clips were recorded at an industrial frame rate of 30 frames per second. In the dataset, the mentor, who is an experienced nurse licensed in Taiwan, performs 11 commonly prescribed physiotherapy movements (i.e., $K = 11$) within a 19.6-second clip (588 frames). The convalescent performs the same set of movements in a separate 20-second clip (601 frames). To evaluate the robustness of matching under different camera placements, we recorded the mentor from a frontal view, while the convalescent was recorded simultaneously from three distinct angles: front, left, and right.

B. Qualitative Study

We applied the proposed method to real-world testing scenarios, which are also used in the subsequent quantitative evaluations. Fig. 4 shows a screenshot from our demonstration video, with full comparisons available at our GitHub repository: https://github.com/NCKU-CIoTlab/TALMA-on-ALPS/. To evaluate the method's robustness under varying viewpoints, we tested three mentor-convalescent camera-angle pairings: front-front, front-left, and front-right. Notably, the convalescent intentionally skipped the eleventh movement. Our system correctly detected this omission by giving a low similarity score, demonstrating its ability to handle incomplete action sequences.

Beyond camera-angle invariance, our system is also designed to cope with imprecise, incomplete, or low-quality movements, which are common in home-based rehabilitation. When a convalescent fails to complete a movement, the system still returns the most semantically similar frame, accompanied by a visibly low similarity score displayed in red text on the user interface. This feature enables physiotherapists to easily identify missed or substandard movements and respond accordingly, significantly enhancing the system's practical applicability in real-world, non-ideal rehabilitation conditions.

C. Precision Study

Table I shows the MAE of various methods under different camera-angle configurations. The settings are represented as mentor-convalescent camera-angle pairs, where F, L, R denote Front, Left, and Right camera positions, respectively. For fairness, all seven evaluated methods share the same overall algorithmic structure, making this experiment functionally equivalent to an ablation study. Our proposed method, shown in the last row of Table I, achieves the lowest overall MAE(1.6) and smallest variance (ranging from 1.2 to 2.1 across all camera settings), demonstrating high precision and robustness. For reference, an MAE of 2.1 corresponds to a temporal deviation of less than $2.1/30 \leq 0.07$ seconds at a frame rate of 30 fps, making the difference nearly imperceptible to human observers.

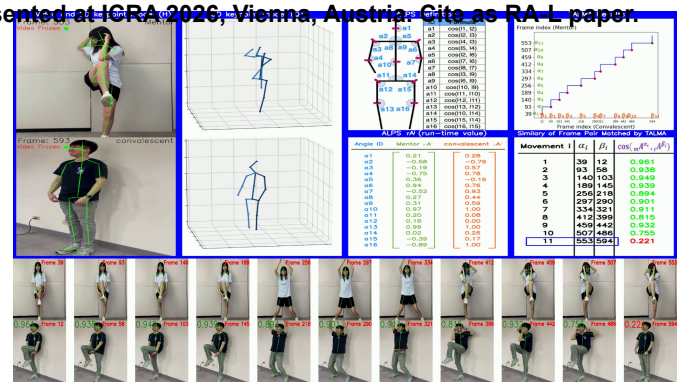


Fig. 4. Screenshot from the demonstration video illustrating the proposed physiotherapy video matching method. The figure visualizes the full pipeline, including 2D/3D pose estimation, limb angle computation, ALPS vector construction, and frame-wise alignment results between mentor and convalescent under different camera angles.

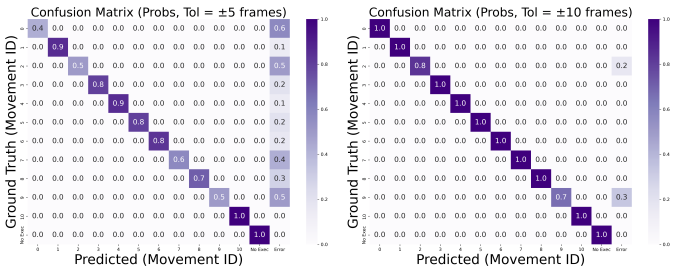
More detailed analysis is described below. The last two rows of Table I highlight the benefit of using three ALPS variants (full-body, left-body, right-body) in Algorithm 1. Specifically, the ^{F}ALPS+TALMA setting (second-to-last row) yields a noticeably higher MAE in the F-R scenario, indicating the advantage of incorporating partial-body ALPS to handle occlusion and asymmetry more effectively. In constrast, the 2D keypoint-based methods (rows 1 and 4-5) produce higher MAEs, even paired with TALMA. These methods are more sensitive to camera-angle variation and often fail to capture subtle postural changes, resulting in lowerquality matches. The instability of 2D methods is particularly evident: for example, in rows 4-5, the MAE is 1.9 in F-F setting but increases sharply to ≥ 8.1 and ≥ 10.8 in F-L and F-R, respectively. These results underscore the stability and viewpoint invariance provided by the ALPS representation.

Rows 1-3 represent the baseline matching performance using DTW [16] without semantic enhancement. Even when DTW is applied to ALPS vectors (rows 2-3), the MAEs are still higher compared to TALMA-based approaches (rows 4-7), reinforcing the importance of incorporating physiotherapy-specific semantics in the matching process. Finally, we conducted a supplementary analysis on the sensitivity of TALMA's hyperparameters θ_{sim} and ϕ used in Step 2.2 of Algorithm 1. Due to space limitations, the detailed findings are presented in our technical report [17].

TABLE I. Comparison of MAE (in frames) across different methods and mentor-convalescent camera-angle configurations. For all methods, 2D keypoints are extracted using Alphapose [5], and DTW results are based on our re-implementation of [16] using the proposed structural definitions.

Method	F-F	F-L	F-R	Overall
^{F}2D-Keypoints+DTW	145.5	105.2	134.9	128.5
^{F}ALPS+DTW	82.3	112.3	31.4	75.3
$^{[F\|L\|R]}$ALPS+DTW	88.9	175.6	87.4	117.3
^{F}2D-Keypoints+TALMA	1.9	8.7	10.9	7.1
$^{[F|L|R]}$2D-Keypoints+TALMA	1.9	8.1	10.8	6.9
^{F}ALPS+TALMA	**1.2**	**2.1**	2.2	1.8
$^{[F\|L\|R]}$ALPS+TALMA	**1.2**	**2.1**	**1.7**	**1.6**

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a). Strict Tol. ± 5 , acc. 0.74 (b). Moderate Tol. ± 10 , acc. 0.96

Fig. 5. Confusion matrices evaluated under two clinically motivated temporal tolerance windows. The prediction-only column `Error` reflects frame-level misalignment, while the bidirectional class `NoExec` highlights cases of undetected or skipped actions, revealing typical mismatches and omissions in physiotherapy execution.

D. Robustness Study Under Challenging Conditions

To evaluate the robustness of our system under realistic and challenging environments, we conducted an experiment simulating common home-use scenarios by placing the camera at ground level, resulting in an upward-facing view. This configuration reflects practical situations where patients may position recording devices on the floor or low furniture, often leading to severe viewpoint distortion compared to standard front-facing setups.

Under this camera angle, we introduced four types of abnormal patient behaviors frequently observed in unsupervised or home-based rehabilitation settings: (1) **Slow-paced execution:** movements are performed at approximately half the mentor's speed; (2) **Lack of actions**: certain prescribed movements are skipped; (3) **Redundant actions**: additional, non-prescribed movements are performed; (4) **Incorrect execution**: movements are deliberately performed with incorrect form.

The upward-facing angle significantly alters the perceived spatial distribution of body keypoints, particularly for the limbs and upper torso, resulting in distorted input features that pose a challenge to conventional pose-matching methods. However, our system leverages a *frame-wise comparison* strategy that does not rely on strict temporal alignment, making it inherently tolerant to temporal irregularities, such as those caused by variable movement speeds.

Table II presents the MAE (in frames) for each abnormal behavior scenario. Using the $^{[F|L|R]}$ALPS+TALMA configuration, our method achieved: 7.3 frames for *Slow-paced execution*, 5.2 frames for *Lack of actions*, 5.8 frames for *Redundant actions*, and 5.4 frames for *Incorrect execution*, with an overall average MAE of **5.9 frames**. While this represents a slight drop in performance compared to standard views, it clearly demonstrates the robustness and practical viability of our approach under both visual distortion and behavioral deviation.

TABLE II. MAE (in frames) under various abnormal behavior scenarios with an upward-facing camera view. The evaluated scenarios include slow-paced execution, missing actions, redundant actions, and incorrect movement execution.

Method	Slow	Lack	Redundant	Incorrect	Overall
$^{[F\|L\|R]}$ALPS+TALMA	7.3	5.2	5.8	5.4	5.9

E. Confusion Matrix Analysis

Figure 5 shows confusion matrices evaluated under two clinically inspired tolerance windows, which fall well within one second, commonly used by physiotherapists for frame-level review. These windows reflect varying degrees of strictness in temporal alignment:

- **Tighter window:** ± 5 frames (≈ 0.17 s) — yielding an overall accuracy of 0.74.
- **Moderate window:** ± 10 frames (≈ 0.33 s) — yielding an overall accuracy of 0.96.

We evaluate a total of seven video scenarios, including both normal and challenging cases:

- **Four standard camera-angle configurations:** front–front, front–left, front–right, and front–low (upward-facing view).
- **Three abnormal scenarios** under the front-low (upward-facing) view, simulating common at-home **rehabilitation issues:** (1) Slow execution (approximately half speed), (2) Missing actions, (3) Extra actions.[4]

For each mentor anchor frame α_i, the system predicts a matching convalescent frame β_i, which is treated as a classification outcome. To better reflect real-world physiotherapy settings, we introduce two additional semantic classes:

- **`NoExec (No Execution)`**: The predicted frame has a similarity score < 0.7, indicating the patient likely did not perform the prescribed action. Although the system still outputs the closest-matching frame, the low score denotes an execution failure. This class appears on both the ground-truth and prediction axes.
- **`Error`**: The predicted frame has a similarity score ≥ 0.7, but the temporal gap $|\alpha_i - \beta_i|$ exceeds the tolerance window. This class, appearing only on the prediction axis, captures high-confidence yet misaligned matches.

Despite these challenging conditions, including severe viewpoint distortions and behavioral inconsistencies, TALMA maintains robustness. Under moderate tolerance windows, the system achieves 96% overall accuracy, underscoring its reliability and adaptability in real-world physiotherapy applications.

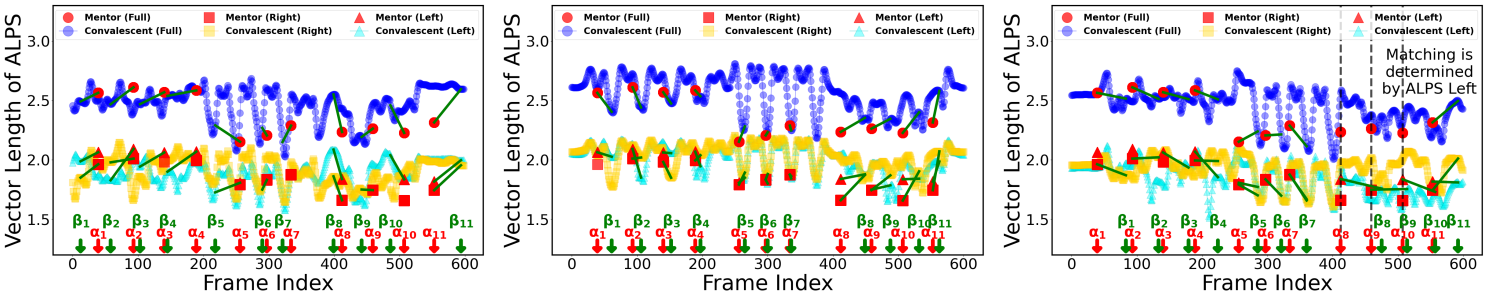
F. Interpretation of TALMA Behavior

Fig. 6 shows the matching behavior of the proposed Three Phase ALPS Matching Algorithm (TALMA) by illustrating the alignment between the mentor's anchor frames and the convalescent video frames. In each subfigure, the horizontal axis represents the frame index, and the vertical axis is the length of an ALPS vector, i.e., mapping the vector $\mathbf{A}_{16\times 1}$ to $\|\mathbf{A}\|_2 = (\mathbf{A}^\top \mathbf{A})^{\frac{1}{2}} \in \mathbb{R}$, to visualize the matching relationship in a two-dimensional plot. The 11 anchor frames $\alpha_1, \ldots \alpha_{11}$ and their corresponding matches $\hat{Z}^* = \{\beta_1, \ldots, \beta_{11}\}$ obtained by Algorithm 1, are indicated by red and green markers, respectively. Each matched pair (α_i, β_i) is connected by a green line, indicating the source ALPS model used in the final decision made by TALMA.

In cases where multiple ALPS models yield the same similarity score to a given anchor frame, the system may identify β_i from more than one model. For example, in Fig.

[4]The erroneous-action case reuses the slow-execution clip and thus does not increase the total number of scenarios beyond seven.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Front-Left (b) Front-Front (c) Front-Right

Fig. 6. Visualization of the Three-Phase ALPS Matching Algorithm (TALMA) in physiotherapy video matching. The matching results are shown for three mentor–convalescent camera-angle configurations: front-front, front-left, and front-right. The vertical dashed lines in Fig. 6(c) highlight specific instances where TALMA, leveraging all three ALPS variants (full-body, left-body, and right-body), yields more accurate matches compared to using full-body ALPS alone.

6(b), the similarity scores between α_1 and β_1 from both full-body ALPS and left-body ALPS are equal (as determined in Line 52 of Algorithm 1, and thus green lines are drawn from both mentor views. This figure effectively illustrates Step 3 of Algorithm 1 – the integration phase, where the system selects the best match among multiple ALPS perspectives.

From Fig. 6(b) to Fig. 6(c), we observe that 30 out of 33 anchor frames in the test clips are successfully matched by using only ^{F}ALPS (full-body ALPS), explaining why row 5 (${}^{[F|L|R]}$ALPS+TALMA) in Table I performs comparably to row 6 (^{F}ALPS+TALMA). The primary distinction between these two methods arises in matching anchor frames α_8-α_{10} as seen in Fig. 6(c) (marked with vertical dashed lines). In these cases, left-body ALPS (^{L}ALPS) provides a higher similarity score than full-body ALPS. This discrepancy can be attributed to the camera placement: The testing video in Fig. 6(c) was recorded from a right-front diagonal angle, which enhances the visibility of the left-side body features. As a result, accurately matching these movements requires a stronger sensitivity to the left limb postures, which our system addresses through the use of three ALPS models. This confirms the method's robustness in matching physiotherapy videos across varied viewpoints and setups.

VI. Conclusions

This study revisits physiotherapy services by enabling convalescents to record exercises at home using mobile devices from arbitrary angles. To support accurate remote assessment, we propose a physiotherapy video matching (PVM) method that handles varied viewpoints. Key contributions include: (1) formalizing PVM as an optimization problem and solving it via a modular pipeline using Alphapose and DST; (2) introducing the Angle-of-Limb-based Posture Structure (ALPS) and Camera-Angle-Free (CAFE) transformation to achieve angle-invariant representation; and (3) developing the Three-Phase ALPS Matching Algorithm (TALMA) for precise movement alignment. Real-world experiments confirm our method's accuracy, robustness, and interpretability. This work offers a foundation for scalable remote rehabilitation. Future directions include reducing latency for real-time use, tuning key parameters in TALMA to improve adaptability, and expanding to multi-person scenarios such as group physiotherapy or team sports analysis.

References

[1] J. S. Bolwell, "A Review of Healthcare Challenges in the UK and the US: Medical Errors, Aging, Private Healthcare and Governance," *Health Sciences Review*, vol. 14, p. 100211, 2025.

[2] W. Kim, "Exploring the Structural Reform of Youth Policies to Promote Fertility," International Center for Public Policy, Andrew Young School of Policy Studies, Georgia State University, Working Paper 2501, 2025.

[3] E. Abbott, A. Campbell, and S. Tidman, "Reliability of Single-Plane Movement Observations in Physiotherapy Assessments," *Rehabilitation and Posture Analysis*, vol. 29, no. 4, pp. 301–315, 2023.

[4] R. Ruivo, P. Pezarat-Correia, and A. Carita, "Static Posture Assessment Using Software Analysis," *Posture Sci.*, vol. 15, no. 4, pp. 202–215, 2023.

[5] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body Regional Multi-person Pose Estimation and Tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2022.

[6] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A Spatio-temporal Transformer for 3D Human Motion Prediction," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 565–574.

[7] A. Z. Skouras, A. K. Kanellopoulos, S. Stasi, A. Triantafyllou, P. Koulouvaris, G. Papagiannis, and G. Papathanasiou, "Clinical Significance of the Static and Dynamic Q-angle," *Cureus*, vol. 14, no. 5, 2022.

[8] P. Kejonen and K. Kauranen, "Reliability and Validity of Motion Analysis for Standing Balance Measurements," *Balance and Posture Analysis*, vol. 22, no. 3, pp. 140–155, 2023.

[9] N. Stergiou and R. Harbourne, "Optimal Movement Variability in Neurologic Physiotherapy," *Neurologic Rehabilitation*, vol. 11, no. 2, pp. 78–92, 2023.

[10] P. Levinger and W. Gilleard, "Postural Assessment and Foot Motion Analysis," *Postural Analysis*, vol. 20, no. 6, pp. 350–365, 2023.

[11] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625.

[12] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriu, L. Romeo, and F. Verdini, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.

[13] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2405–2414.

[14] E. Ronn, "NP-complete stable matching problems," *Journal of Algorithms*, vol. 11, no. 2, pp. 285–304, 1990.

[15] C.-C. Chen, M.-H. Hung, B. Suryajaya, Y.-C. Lin, H.-C. Yang, H.-C. Huang, and F.-T. Cheng, "A Novel Efficient Big Data Processing Scheme for Feature Extraction in Electrical Discharge Machining," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 910–917, 2019.

[16] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[17] J.-W. Lin, P.-W. Chen, Y.-P. Huang, M.-H. Hung, M.-H. Kao, J. Ji, L.-Y. Jiang, Y.-S. Chou, and C.-C. Chen, "A Physiotherapy Video Matching Method Supporting Arbitrary Camera Placement via Angle-of-Limb-based Posture Structures," *Technical Report, Available at https://github.com/NCKU-CIoTlab/TALMA-on-ALPS/*, 2025.