Automatic Evaluation of Sports Motion: A Generic Computation of Spatial and Temporal Errors

Marion Morel[a,b], Catherine Achard[a], Richard Kulpa[b], Séverine Dubuisson[a]

[a] *Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7222, ISIR, F-75005, Paris, France*
[b] *M2S laboratory, Université Rennes 2, ENS Rennes, Avenue Robert Schuman, 35170 Bruz, France*

Abstract

In this paper, we propose an innovative automatic evaluation process for any sport motions. Based on a 2-level Dynamic Time Warping, the process allows the evaluation of both spatial and temporal errors of a novice motion based on an experts' motion database and without any prior knowledge on the sport. This new methodology is evaluated with regards to coaches' assessment on two different kinds of motions: tennis serves and karate *tsuki*.

Keywords: Dynamic Time Warping, Evaluation, Multidimensional features, Synchrony, Motion Capture

1. Introduction

Motion sensors are part of our lives in devices such as smartphones or gamepads. They give information about the motion of the device and thus indirectly of the user body part that holds it. Moreover in recent years, the emergence of new low-cost motion capture systems such as Microsoft Kinect or Noitom Perception Neuron allows access to full-body motion. All these devices are thus radically changing the interaction users have with computers or consoles. It will even lead to a new generation of sports training tools based on the motions of the subject. However, the system must be able to evaluate its performance, to identify its errors and then to propose a way to correct them. This evaluation is a challenging task due to the large types of sports and the variability in morphology and style of the subjects.

In this paper, we propose a new generic motion evaluation method that automatically identifies the spatial and temporal errors performed by the subject, whatever the sport. It is only based on a database of motions performed by experts that includes the different correct ways to perform the motion.

To our knowledge, sport gestures were never studied this way, namely (i) **without any prior knowledge** on the motion to be executed, except some instances made by experts, (ii) by estimating their qualities limb by limb, (iii) by detecting their errors to be corrected. Some approaches exist where an expert specifies what is a good gesture (for a karate kata, the kicking wrist must have a linear trajectory for example). Then, the system has to determine if a new gesture checks or not these rules and possibly evaluates the distance from these rules. The approach proposed in this paper strongly differs from the previous ones in the sense that it is only based on a set of expert gestures. The system learns from this set what is important and has to be checked,

for each time of the gesture and each limb. The advantage of this approach is twofold: first, it can be generalized to all gestures provided that instances are available. Second, it allows to automatically determine the errors.

There is a similar challenge in the field of surgery as presented in the review [26]. Although the required skill for surgeons is strongly different, the objective is the same: achieving a perfect gesture to improve the surgery task. In this specific case, the approaches focus on the tool used by the surgeon: this makes the analyses very different. Typically, in laparoscopic surgery, Hofstad *et al.* [7] evaluated gesture performance through the softness of the motion of the tool, its average speed and its global trajectory.

After summarizing related works in Section 2, we will detail our evaluation process in Section 3, that includes three stages: motion encoding, creation of the model of experts' motions and evaluation of a novice's motion. Results will be presented in Section 4. Section 5 concludes this paper and presents some future works.

2. Related Work

Evaluating a motion is complex since it deals with large amounts of data varying in space and time. Thus, assessing the quality of an action requires to determine the kinematic factors that represent the right performance. For this reason, some authors proposed to **add prior knowledge** on the motion to analyze. For example, Burns *et al.* [4] defined a set of kinematic rules that defines how a karate *kata* must be performed, such as the linear trajectory of the kicking wrist. In a similar way, Komura *et al.* [11] used on martial art: (1) the minimization of the global motion's energy during a defense, assuming that the least the subject moves, the easier it is for him to counterattack; (2) the unpredictability of an attack, expressed by the minimization of the motion perpendicular to the frontal vector; or (3) the wrist speed. Ward [33] used several intersegmental angles to compare classical ballet techniques (thoracic anterior/posterior tilt, pelvis obliquity, ankle external rotation, *etc.*). Overall, the results of this kind of studies are very interesting but can only be applied to specific gestures whose main features are *a priori* identified. Our goal is different since we aim at proposing a generic evaluation method that can automatically, and for any sport, determine the important features of expert's motions that are then used to evaluate the performance of a new subject.

Instead of adding prior knowledge on the motion, several authors have worked on the **extraction** of relevant features. Ofli *et al.* and Pazhoumand *et al.* have computed the variance [19] and the entropy [21] of each joint to discriminate the most informative features characterizing the motion and used these criteria to recognize actions. The problem of such approaches is that they loose information about the temporality of the motion. Moreover, Ofli *et al.* assume that the more a joint moves, the more significant is this joint to characterize the gesture. However, this assumption can be irrelevant for some sports such as the hips that should not move during a *tsuki* motion. The significance of a joint is thus dependent on the sport and must then be identified. To this end and to keep only informative data, Ramakrishnan *et al.* [23] and Raptis *et al.* [25] used an angular representation of the motion and applied a principal component analysis to decrease the dimension of the data. This process increases its efficiency by only keeping the information that is useful to segment the motion but it then cannot be used to find the motion errors because all joints are mixed during the encoding. Another approach consists in using a database of experts' motions. Jiang *et al.* [8] applied a SVD (Singular Value Decomposition) on such a set of experts' motions. They then added a new motion to the previous set and assumed that the SVD would strongly change if the motion was different from initial ones. The same idea of adding a new

motion to observe representation changes compared to an initial set has also been used by Barbic *et al.* in [2] to segment motions. These methods based on dimension reduction can thus be used to discriminate motions or to measure their global quality compared to a set of motions but they cannot determine on which joints spatial and temporal errors occurred. The problem is indeed complex since the correct method should keep enough details in the encoding to discriminate a correct motion from another one while removing the variabilities due to different viewpoints, morphologies of subjects or global execution speed.

Some authors worked on these **variabilities**. While *2D* video sequences have to deal with the challenge of viewpoint changes [24, 20], it can be handled with data acquired from motion capture by aligning the coordinate system on a local reference frame linked to the human at each frame [17]. To deal with the morphology, Kulpa *et al.* [12] have proposed some invariant features then used by Sorel *et al.* [30] several years later to recognize an action. Sie *et al.* [29] introduced a simple normalization process, according to the torso length.
The execution speed also can vary from one subject to the other. Moreover, some moments of the motion can be executed faster or slower depending on the individual. Indeed, the relative timing of different limbs contains many information about the motion and is very important to estimate motion quality. Temporal information is thus essential at a global level, considering both the whole body but also some of its local parts, especially the links between limbs. The temporality between motions has been studied in [13] to evaluate dance step quality. The analysis was based on the tempo between music and motion. This application is restricted to the dance, since it manages the synchronization of the motion with an external and specific tempo. In this paper, we also want to evaluate the temporal errors made by body parts but not based on an external source but on a set of experts' motions. The problem is then much harder since the speed profile of each limb can be different, can evolve simultaneously or with a shift for instance and these differences are important to be evaluated since they are the core of the right performance for the motion, even if some variabilities can be found in the global execution speed for instance.

To manage motions having different lengths, speeds and/or different rhythms, some authors used Hidden Markov Models (HMM) or Hidden Conditional Random Field (HCRF) [34, 9, 30, 32] in which states represent postures of motion and transitions between them are defined by probabilities. This model, well-known in temporal pattern recognition, is thus robust to temporal variations. However, several time steps are associated with the same state and the temporality is only managed between these states. The evaluation is then not precise enough.

To evaluate the **synchrony** (or temporal coordination) between different limbs, we need more accurate methods that study the alignment between temporal signals. Dynamic Time Warping (DTW) [27], originally developed for speech processing, is a well-established method to take into account the temporal variations for the comparison of time series. Many studies have tried to improve the efficiency of the DTW algorithm over recent years depending on the context of applications [10, 36, 35, 6, 5].

In the context of gesture, DTW has been used by several authors to align motions and to compare them. Pham *et al.* [22] computed invariant curvatures of a surgical tool and then matched it to a template using DTW. This method has been applied in obstetrics to compare forceps blade placements between a senior medical doctor and a novice. Sakurai *et al.* [28] proposed a method to retrieve a gesture similar to the performed one from a database. The system normalized the new gesture captured by the Kinect and computed the area covered by the skeleton to be employed as feature parameter. A DTW was then used to measure the similarity between the new motion and those in the database. In these studies, the whole motion was encoded using a single temporal series (from only one body part) and thus no information was

kept on the different limbs or their temporal coordination (synchrony). This encoding did not allow a fine analysis of motion and an error detection during their execution. To our knowledge, only Morel *et al.* compared local and global alignments of two skeletons performing the same gesture in [17]. In this previous work, a global alignment is first done on the whole body to temporally readjust both gestures. Then, the spatial errors are estimated for each joint that have been re-aligned with this global path. To estimate if some joints are delayed, a local DTW is done, joint by joint and the alignment computed for each joint is compared to the whole body alignment. This first work has a main drawback: if a gesture is correct, except a limb that is delayed, this delay produces a change of the global alignment that aligns at best all joints. This is not correct because it induces a spatial error for all joints, errors that do not exist. The method proposed in this paper is strongly different since each limb is locally aligned to the expert motion to estimate spatial errors. This allows to avoid the problem previously mentioned. To study the delay, we then just compare the local paths between limbs.

By using DTW or Transported Square-Root Vector Fields (TSRVFs), Veeraraghavan *et al.* [31] and Ben Amor *et al.* [3] used Kendall's shape manifold and introduced the function space of time warping, that models and learns the variability due to execution speed. These works are relevant for action recognition but cannot manage a difference of synchrony between limbs that is different from the global speed of the motion and thus cannot evaluate local temporal errors.

In this paper, we propose an approach to concurrently take temporal and spatial information into account to evaluate motion quality. This approach is generic so it can be applied to any motion and restricts the influence of subject's morphologies. It is based on several DTWs computed globally on the whole motion and locally on each limb. This allows the extraction of spatial and temporal information to determine, for incorrectly performed motions, at which time and why they are badly executed.

Temporal and spatial errors are separately estimated to provide insightful and discriminative information about the performance of the player. However, sometimes a player can try to compensate a spatial error by changing the timing of his gesture and inversely. Our method does not consider this relation, the strategy or the cause of the change and only observes the results on the motion performed. These changes indeed are not harmless. For example, the modification of speed to correct a spatial error leads to an important increase of joint forces and thus of risk of injury [16]. Even some slight temporal changes in the motion can lead to a different segmental sequence and a modification of the energy flow [15]. These motions cannot thus be considered as correct ones. For all these reasons, we tried to evaluate the temporal and spatial errors independently without taking care of the source of these errors. The genericity of the method is obtained thanks to an experts' motion database from which a nominal motion is extracted as well as the allowed deviations around this motion for each limb.

3. Methodology

The goal of our work is to assess the quality of any sport motion given a set of experts' motions. This comparison of motions is complex since subjects may have different morphologies and may perform the motion at different speeds. Moreover, spatial variabilities may be due to a specific execution style or to a bad execution of the motion. To deal with all these variabilities, let us consider a set of experts' motions including as much interpersonal variability as possible. This set is used to learn a representative expert motion that we call *nominal motion* (following the terminology in [31]) and a spatial tolerance around this nominal motion, for each limb.

3.1. Notations and gesture encoding

To create the database, the gestures were captured with a `Vicon MX-40` optical motion capture system (Oxford Metrics Inc., Oxford, UK). The subjects were equipped with 43 reflective markers placed on anatomical landmarks to compute the trajectories of the 25 joint centers. A normalization process was carried out to make all motions invariant to the initial position and orientation. All postures are first centered on the same root position and are then oriented so the subject's hips are aligned (Figure 1 describes the location of joints and root). Moreover, to decrease the influence of the morphology, each joint coordinate is divided by the distance between head and root joints as proposed in [29].

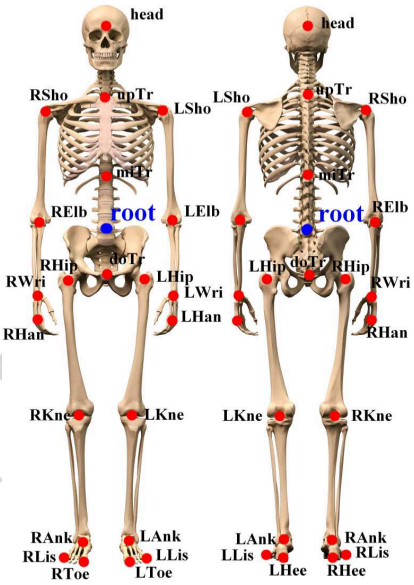


Figure 1: Position of the root and joints of the skeleton

Let us now consider the following notations:

- J: number of joints (25 here).
- L: number of limbs (5 here: right arm, left arm, right leg, left leg, trunk).
- M_i: number of time steps of the i^{th} motion.
- N_E and N_N: respectively numbers of expert motions and novice motions.
- $\mathbf{X}_i(t) = \{\mathbf{x}_i^j(t), j = 1...J\}$, with
  $\mathbf{x}_i^j(t) = (x_i^j(t), y_i^j(t), z_i^j(t))$ the *3D* trajectory of the j^{th} joint and the i^{th} motion.
  Thus, $\mathbf{X}_i(t)$ is a 75-dimensional vector (25×3) that encodes, at time t, the body posture (position of the 25 joints) while $\mathbf{x}_i^j(t)$, $j = \{1...J\}$ only encodes the position of joint j at time t for the i^{th} motion (*3D* vector).

- $\mathbf{X}_i^l(t) = \{\mathbf{x}_i^j(t), j \in S_l\}$ where $S_l = \{j \in \{1...J\} \mid j$ belongs to limb $l\}$ encodes the position of joints $j \in S_l$ belonging to the lth limb, at time t.

3.2. Model of experts' motion

All experts performed the motion accurately as in real situation so their execution speeds may be different. To take into account these speeds variations that can be nonlinear, all experts' motions were aligned with a Dynamic Time Warping (DTW) method that is briefly described below.

3.2.1. Dynamic Time Warping

Let us first consider two motions $\mathbf{X}_0(t)$ and $\mathbf{X}_1(t')$ and define a distance matrix with elements $d_{0,1}(t,t')$ that represents the distance between $\mathbf{X}_0(t)$ and $\mathbf{X}_1(t')$, $\forall t \in [1, M_0]$ and $\forall t' \in [1, M_1]$ where M_0 and M_1 are the durations of trajectories $\mathbf{X}_0(t)$ and $\mathbf{X}_1(t')$ respectively.

As an Euclidean distance would be very sensitive to the amplitude of both signals [6], we rather chose to compute $d_{0,1}(t,t')$ using both joint position and velocity as suggested in [10]:

$$d_{0,1}(t,t') = \alpha\|\mathbf{X}_0(t) - \mathbf{X}_1(t')\|^2 + \beta\|\dot{\mathbf{X}}_0(t) - \dot{\mathbf{X}}_1(t')\|^2 \tag{1}$$

where α and β are weighting coefficients that give equal influence to trajectories and velocities (empirically determined).

The optimal warping curve $\phi(k)$, k=*1...K* remaps the indices of $\mathbf{X}_0(t)$ and $\mathbf{X}_1(t')$ with the lowest cumulative distance. K that represents the length of the warping curve can vary depending on the mapping of the two motions $\mathbf{X}_0(t)$ and $\mathbf{X}_1(t')$.

$$\begin{aligned}
&\phi(k) = (\phi_0(k), \phi_1(k)) \text{ with}\\
&\phi_0(k) \in \{1...M_0\}\\
&\phi_1(k) \in \{1...M_1\}
\end{aligned}$$

$$\phi = \underset{\phi}{\operatorname{argmin}} \sum_{k=1}^{K} d_{0,1}(\phi_0(k), \phi_1(k)) \tag{2}$$

Monotonicity on ϕ_0 and ϕ_1 is required to guarantee time ordering. Similarly, boundary conditions ($\phi_0(1) = \phi_1(1) = 0$, $\phi_0(K) = M_0$ and $\phi_1(K) = M_1$) and step size conditions ($0 \le \phi_0(k) - \phi_0(k-1) \le 1$ and $0 \le \phi_1(k) - \phi_1(k-1) \le 1 \quad \forall k \in \{2...K-1\}$) are also required.

Let us define $\phi_{1\to0}(t)$ and $\phi_{0\to1}(t')$ the compressed warping functions that project the warping functions on the reference time of $\mathbf{X}_0(t)$ or $\mathbf{X}_1(t')$, respectively:

$$\phi_{1\to0}(t) = \min_{k|\phi_0(k)=t} \{\phi_1(k)\} \tag{3}$$

$$\phi_{0\to1}(t') = \min_{k|\phi_1(k)=t'} \{\phi_0(k)\} \tag{4}$$

Then $P_{1\to0} = \{(t, \phi_{1\to0}(t)) \mid t = 1...M_0\}$ is the warping function that maps $\mathbf{X}_1(t')$ on $\mathbf{X}_0(t)$ whereas $P_{0\to1} = \{(t', \phi_{0\to1}(t')) \mid t' = 1...M_1\}$ is the warping function that maps $\mathbf{X}_0(t)$ on $\mathbf{X}_1(t')$.

3.2.2. Nominal Motion

Let us consider (without loss of generality) that $\mathbf{X}_0(t)$ is the longest expert's motion. The other expert motions $\mathbf{X}_i(t)$ are realigned to $\mathbf{X}_0(t)$ according to the path $P_{i\to 0}$ to obtain new motions $\mathbf{X}_i(\phi_{i\to 0}(t))$ with the same duration of M_0 time steps. $\mathbf{X}_0(t)$ is chosen as the longest expert's motion to reduce the potential loss of information introduced by the alignment process. The nominal motion is then simply computed as the average of all aligned experts' motions:

$$\mathbf{X}_n(t) = \frac{1}{N_E} \sum_{i \in experts} \mathbf{X}_i(\phi_{i\to 0}(t)) \quad \forall t \in \{1...M_0\} \tag{5}$$

The temporal alignment between trajectories is thus the same for the whole body, ensuring to maintain the temporal coherency between joints while being able to obtain the average trajectory of all joints: $\mathbf{X}_n(t) = \{\mathbf{x}_n^j(t), j = 1...J\}$. Figure 2a illustrates the extraction of a nominal pose from all aligned experts' motions.

Note that according to its computation process, the skeleton of the nominal motion $\mathbf{X}_n(t)$ does not preserve the morphological structure of the body. It must thus be considered as a mathematical tool and not a real motion.

In addition to this nominal motion that represents all experts' motions, we need to model the spatial tolerance (deviation) that encodes the allowed variability around the nominal motion.

3.2.3. Spatial tolerance

To be generic, our method must be able to determine the allowed spatial variabilities for each joint around the nominal motion. The quality of a new motion indeed depends on the spatial tolerance experts have for each joint, according to the type of motion. For example, the amplitude of the hip is smaller for all experts than the one of the active wrist during a tennis serve. The spatial tolerance should thus be smaller for the hip than for the wrist and a similar amplitude in the novice motion for these joints can exhibit an error for the hip while it is not for the wrist.

A first idea could be to compute the variability of each joint when the motions are aligned on the nominal motion: $\Sigma^j_{S\,pa}(t) = COV\{\mathbf{x}_i^j(\phi_{i\to n}(t))\}_{i\in experts}$ where $\phi_{i\to n}$ is such that $P_{i\to n} = \{t, \phi_{i\to n}(t), t = 1...M_0\}$ corresponds to the warping function that maps $\mathbf{X}_i(t)$ on the nominal motion $\mathbf{X}_n(t)$.

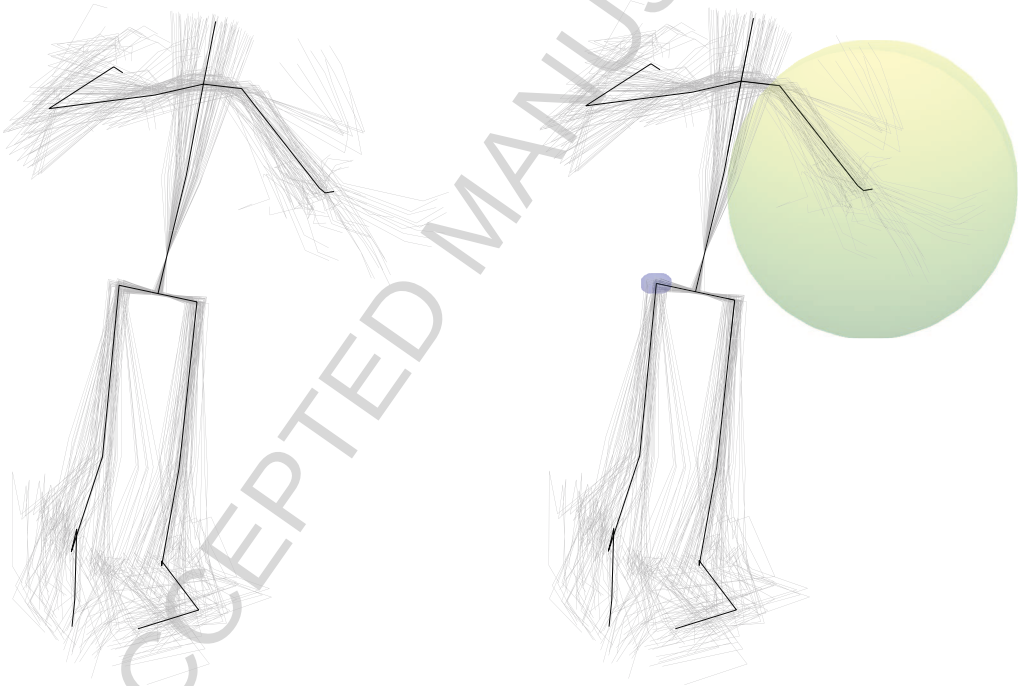
Even if this global alignment $P_{i\to n}$ is the best warping function for the whole body motion, it is not necessarily the best one for each limb independently. To be more accurate in identifying local errors, we introduce local alignments based on the sets of joints composing each limb l, called S_l. Let $P^l_{i\to n}$ be the warping function that maps $\mathbf{X}^l_i(t)$ on $\mathbf{X}^l_n(t)$ (expert's and nominal motions restricted to the l^{th} limb):

$$P^l_{i\to n} = \{t, \phi^l_{i\to n}(t) \mid t = 1...M_0\} \tag{6}$$

From the local alignment we define the spatial tolerance of the l^{th} limb as the variability of experts' positions around the nominal one at a any time t:

$$\Sigma^l_{S\,pa}(t) \quad = \quad COV\{\mathbf{X}^l_i(\phi^l_{i\to n}(t))\}_{i\in experts} \tag{7}$$

where COV is the covariance matrix of the vector containing the concatenation of all $3D$ positions of each joint of S_l at time t. To reduce the computational complexity, we impose the covariance



(a) A nominal motion's pose (in black) extracted from the experts' aligned motions (in gray).

(b) Spatial tolerance of the left wrist (in green-yellow) and the right hip (in blue).
(Best viewed in color)

Figure 2: Nominal motion and spatial tolerances

matrix to be block diagonal (then the covariance is only computed on same joints coordinates). Figure 2b depicts the spatial tolerance of the left wrist and the right hip at a specific time step. As expected, the spatial tolerance of the left wrist is much larger than the right hip's one.

3.3. Evaluation of a novice's motion

Let us consider a novice motion $\mathbf{X}_i(t)$. To estimate the performance, a qualitative evaluation must be carried out according to two criteria. The first criterion is linked to the spatial error each limb has at any time of the motion. It helps identifying a joint that is not moved correctly for instance. However, even if the spatial performance of all joints is performed correctly, the motion can be wrong if the coordination between these joints is not correct. For example, a correct tennis serve must exhibit a right segmental sequence from the legs to the hitting hand. The second criterion is thus the temporal error between a pair of limbs. This section describes the evaluation of both spatial and temporal errors for each limb and each time step.

3.3.1. Spatial Errors

Spatial errors are estimated for each time t and each limb l, relatively to the spatial tolerance allowed around the nominal motion. The trajectory of each limb l of the novice motion is first aligned on the trajectory of the l^{th} limb of the nominal motion. Then, the spatial error of limb l is given by the Mahalanobis distance between both signals $\mathbf{X}_i^l(\phi_{i\to n}^l(t))$ and $\mathbf{X}_n^l(t)$:

$$E_{S\,pa,i}^l(t) = \sqrt{F_i^l(t)^T \left(\Sigma_{S\,pa}^l(t)\right)^{-1} F_i^l(t)} \quad \forall t \in \{1...M_0\} \tag{8}$$

with:

$$F_i^l(t) = \mathbf{X}_i^l\left(\phi_{i\to n}^l(t)\right) - \mathbf{X}_n^l(t) \tag{9}$$

3.3.2. Temporal Errors

Let l_1 and l_2 be the two limbs we want to compute the temporal error on:

- $P_{i\to n}^{l_1} = \{t, \phi_{i\to n}^{l_1}(t) \mid t = 1...M_0\}$ is the local alignment of limb l_1 between a novice and the nominal motion.
- $P_{i\to n}^{l_2} = \{t, \phi_{i\to n}^{l_2}(t) \mid t = 1...M_0\}$ is the local alignment of limb l_2 between a novice and the nominal motion.

If limbs l_1 and l_2 have the same coordination for novice and nominal gestures, then the warping function based on the restricted limb l_1 would be the same than the warping function based on l_2: $\phi_{i\to n}^{l_1}(t) = \phi_{i\to n}^{l_2}(t), \quad \forall t \in \{1...M_0\}$. Thus, we define the temporal lag of l_1 relatively to l_2 at time step t as the delay between $\phi_{i\to n}^{l_1}$ and $\phi_{i\to n}^{l_2}$:

$$E_i^{l_1,l_2}(t) = \phi_{i\to n}^{l_2}(t) - \phi_{i\to n}^{l_1}(t) \quad \forall t \in \{1...M_0\} \tag{10}$$

However, this temporal lag is only meaningful if limbs are moving. Actually, for static limbs, the synchronization made by the warping functions can be not significant. We thus introduce weighting coefficients $\gamma^{l_1,l_2}(t)$ that give more influence to the time steps with higher speed:

$$E^{l_1,l_2}_{Temp,i}(t) = \gamma^{l_1,l_2}(t) \times \left(\phi^{l_2}_{i\to n}(t) - \phi^{l_1}_{i\to n}(t)\right) \tag{11}$$

where

$$\gamma^{l_1,l_2}(t) = \frac{\max(||\dot{\mathbf{X}}_n^{S_{l_1}}(t)||, ||\dot{\mathbf{X}}_n^{S_{l_2}}(t)||)}{\sum_{t=1}^{M_0} \max(||\dot{\mathbf{X}}_n^{S_{l_1}}(t)||, ||\dot{\mathbf{X}}_n^{S_{l_2}}(t)||)} \tag{12}$$

and $||\dot{X}(t)||$ is the magnitude of the velocity of $X(t)$.

4. Results

4.1. Datasets and annotations

Two datasets were used to validate our work. They were both partially annotated. The first one contains tennis serves and allows the assessment of our spatial evaluation. However, as described below, the experts' evaluation of tennis serves is mainly based on postures so we used a second dataset of karate motions that is used to evaluate both our temporal and spatial errors. Let us briefly describe these 2 datasets.

Tennis serve dataset. The dataset is composed of 75 experts' tennis serves performed by 9 national ranked players and 72 novices' ones performed by 8 beginners or middle level players. We developed an annotation tool (see Figure 3) and asked a tennis coach to annotate the captured motions (one for each subject). The annotator had to first subdivide the tennis serves into 4 phases according to [14]. Then, he had to provide 20 scores, one for each of the 5 limbs and for each phase of the motion. The annotation tool displays the original motion before the normalization process and joints extraction. It also allows the control of the viewpoint to better analyse the gesture, and a slider can be shifted to control the temporal axis. The racket was added on the annotator's request.

Karate tsuki dataset. The second dataset is composed of 30 experts' and 65 novices' karate motions called *tsuki* performed by 6 experts and 9 novices. This motion has been chosen since it requires a high temporal coordination between both arms. The annotations of *tsuki* motions were done by 2 karate coaches who evaluated the performance of both arms of each subject. The spatial annotation was related to the spatial correctness of both arm trajectories. The temporal annotation was linked to the synchrony between arms. Since the *tsuki* motion is very simple and has only one phase, this annotation is the quantification of the global lag between arms. It varies between -10 (important delay) and 10 (important advance).

The following sections show how these datasets were used to validate the global alignment used to create the nominal motion, then to validate both spatial and temporal errors.

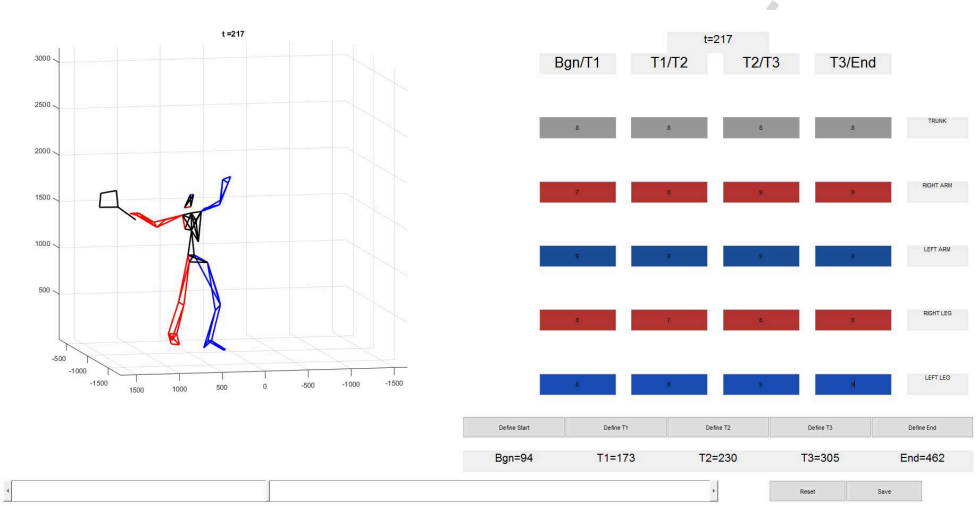


Figure 3: Annotation tool for tennis serve evaluation

4.2. *Validation of nominal motion*

The tennis coach first subdivided each serve trial into the 4 phases defined in [14] from 3 key times: T_1 (racket at highest position), T_2 (racket at lowest position) and T_3 (ball impact). These key times are illustrated on Figure 4. Let us denote $T_{Ann,i}(p)$ the annotation of the p^{th} key time ($\forall p \in \{1...3\}$) and the i^{th} expert motion ($\forall i \in$ experts).
The goal is to assess $\hat{T}_j(k)$ that defines the phases of the motion of novice k, knowing $T_{Ann,i}(p)$, $\forall i \in$ experts, $\forall p \in \{1...3\}$.

To validate the temporal alignment used to create the nominal motion, we made a 2-step process (see Figure 5): we first estimated the phases of the nominal motion (called nominal phases) from the annotations on experts' motions; we then used these nominal phases to estimate the phases of the novice's motions that we compared to the annotated ones.

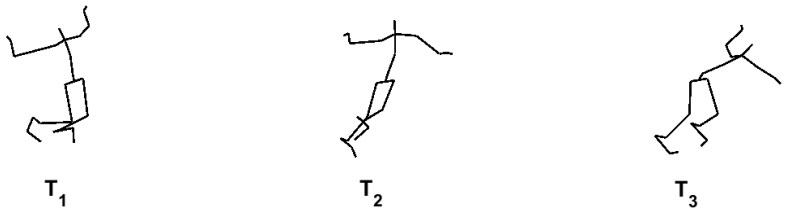
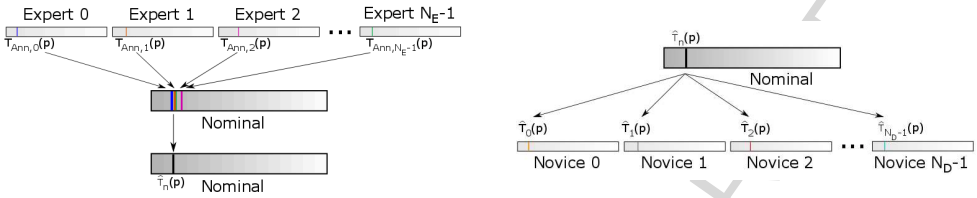


Figure 4: Key times for a tennis serve

Nominal phases estimation
We first used $P_{n\to i} = \{(t, \phi_{n\to i}(t)), t = 1...M_i\}$ to transfer the phase annotations $T_{Ann,i}(p)$ of expert



(a) First step: estimation of nominal phases

(b) Second step: estimation of novice's phases

Figure 5: Motions phases detection

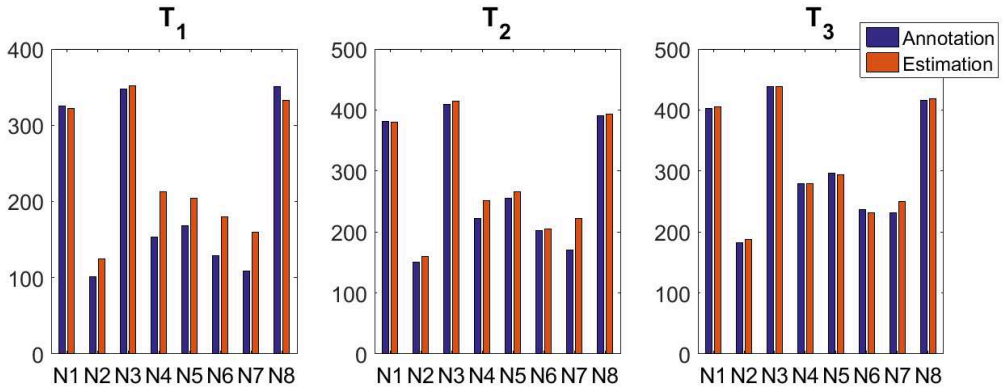


Figure 6: Comparison of the typical key times assessment and annotations

serve i on the temporal axis of the nominal motion:

$$\hat{T}_{n,i}(p) = \phi_{n \to i}(T_{Ann,i}(p)) \tag{13}$$

The nominal phases $\hat{T}_n(p)$ were then estimated as the average of all assessments $\hat{T}_{n,i}(p)$:

$$\hat{T}_n(p) = \frac{1}{N_E} \sum_{i=0}^{N_E-1} \hat{T}_{n,i}(p) \tag{14}$$

Detection of the novices' phases

The key times of novice's motions can then be automatically determined by first globally aligning them with the nominal motion and then by using the warping curve on nominal phases:

$$\hat{T}_k(p) = \phi_{k \to n}(\hat{T}_n(p)) \tag{15}$$

Figure 6 gives a comparison between the results obtained with this process and the coach's annotations, for each of the 8 novices. These estimations are very close to the ground truth of the annotator, even for low quality serve (see the following section).

4.3. Validation of the spatial error

Spatial errors were validated for both datasets as annotations were available.

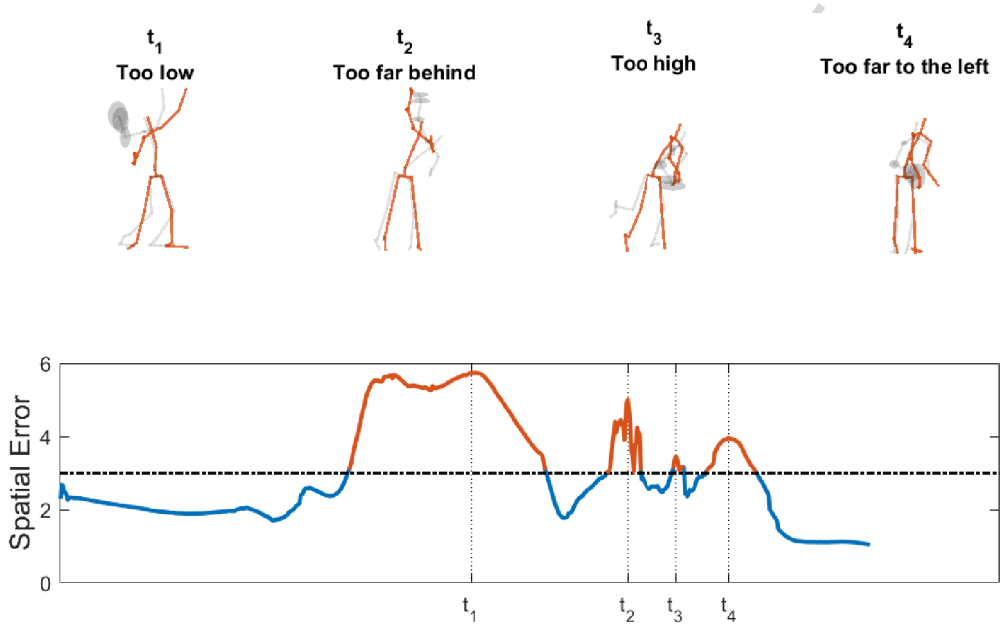


Figure 7: Automatic spatial error detection. First row: red and gray skeletons depict respectively the novice's motion and the nominal one. Gray ellipsoids correspond to the spatial tolerance of each joint of the right arm. Second row: Spatial error of the novice's right arm over time. The four reference times illustrated on the top are represented by vertical dashed lines in the bottom graph. They correspond to frames for which the error is maximal among the range of frames admitting a spatial error exceeding the symbolic value of 3 (threshold such that 99.7% of experts' gestures check the condition considering a gaussian distribution).

4.3.1. Tennis serves

The evaluation process provides a spatial error for each limb at each time step as shown in Figure 7 for the right arm. In this example, the system automatically detects four keys moments corresponding to four spatial errors that are illustrated on top of the figure (focus on the right arm). The interpretation of the errors is simply made by geometric tests. Even if these illustrations are very pertinent to design a virtual coach, they are not pertinent enough to validate the proposed method. Thus, we present a quantitative evaluation based on annotations.

Let us denote $S^l_{Ann,i}(p)$ the coach's annotation of each motion i, for each motion's phase p and each limb l. $S^l_{Ann,i}(p)$ was based on a scale varying from 0 to 10, the higher the better. It can be compared to the average of our spatial error in the corresponding phase:

$$S^l_{Us,i}(p) = \frac{1}{|T_P|} \sum_{t \in T_P} E^l_{Spa,i}(t) \tag{16}$$

with T_P the time range of p^{th} phase.

To present synthetic results, the scores of all limbs have been averaged for each phase of the serve ($\frac{1}{L}\sum_{l=1}^{L} S^l_{Ann,i}(p)$ and $\frac{1}{L}\sum_{l=1}^{L} S^l_{Us,i}(p)$). Thus, four spatial measures (one per phase) are obtained for each motion.

The results are illustrated in Figure 8. A leave-one-subject-out cross validation was carried out on experts' motions: for each annotated motion we tested, the whole set of experts' motions except those performed by the considered subject were used to model the expert's motion. To evaluate the spatial errors of novice's motions, the whole set of experts' motions was used since the novices were not in that set. The annotator's score belongs to interval $[0, 10]$ (0: bad quality, 10: high quality), whereas our errors belong to interval $[+\infty, 0]$, decreasing with the quality of the motion. Given the appearance of the score distribution, an exponential curve best represented the relation between our spatial errors and the annotator's evaluation. The coefficient of correlation of the logarithm of the data is $r = -0,7685$ ($p < 0.05$) and thus shows that there is an exponential relation between the evolution of our spatial errors and the scores made by expert annotator, our ground truth. By the way, a small spatial error has been estimated for all experts in our dataset by both our automatic method and the annotator.

4.3.2. Tsuki motions

To strengthen the validation of our spatial evaluation and check its genericity, we applied our algorithm on a second kind of motion: a karate punch called *tsuki*.

2 karate coaches provided global spatial scores for each of the 14 performers i, denoted $S^1_{Ann_1,i}$ and $S^1_{Ann_2,i}$ (right arm) and $S^2_{Ann_1,i}$ and $S^2_{Ann_2,i}$ (left arm). As the motion was not subdivided into phases, we simply compared $\frac{1}{2}\sum_{l=1}^{2} S^l_{Ann_1,i}$ or $\frac{1}{2}\sum_{l=1}^{2} S^l_{Ann_2,i}$ to the score provided by our algorithm $\frac{1}{2}\sum_{l=1}^{2} S^l_{Us,i} = \frac{1}{2}\sum_{l=1}^{2}\left(\frac{1}{M_0}\sum_{t=1}^{M_0} E^l_{Spa,i}(t)\right) \quad \forall i \in \{1...14\}$. As for tennis serves, the leave-one-subject-out cross validation was used to evaluate the experts' motions. The results are depicted in Figure 9. Once again, an exponential curve can map our spatial errors to the annotator scores.

Coefficients of correlation applied to the logarithm of the data are relatively high in terms of amplitude: $r = -0.6080$ ($p < 0.05$) for the first annotator and $r = -0.7370$ ($p < 0.05$) for the second one. However, one can also notice the high variability among the coaches' annotations. Despite an overall consensus on the subjects' ranking, the scores' variability (N_4 gets $1.5/10$ and $5.5/10$ for example) makes the fitting sometimes too high, sometimes too low (see N_2, N_4, N_7 and E_3). Our algorithm could thus provide a kind of standardization of the different possible spatial annotations. A larger number of coaches should then be required to find the mean and variability of annotations.

Moreover, the coefficients of correlation were largely decreased (in amplitude) by the N_6 subject evaluation. The coaches and our algorithm both evaluated it as a bad performance but our algorithm exhibited a very high error. It is due to the fact that the subject badly positioned his arms at motion's end (the left fist moves back to the shoulder instead of near the hip). Figure 10 shows two *tsuki* motions performed by an expert (first row) and the novice N_6. Motions were not aligned but just uniformly sampled from the beginning t_1 to the end t_7. It is noticeable that the left arm of novice N_6 is not well positioned at the end of the motion. This very bad location induces a very large spatial error.

Figure 9 also shows in dashed line the equivalent fitting curve obtained for tennis serves. Even with a different sport, we can see that their shapes are similar and even nearly equals between the tennis coach and the first karate annotator, supporting the idea of an automatic motion evaluator whatever the sport or the player.

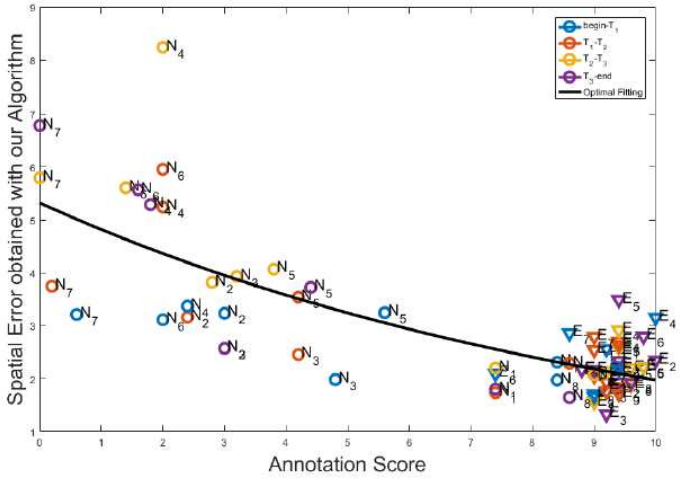


Figure 8: Comparison of spatial error estimations and score annotations for each phase of the tennis serves. Triangles are for experts and circles for novices. The black line represents the exponential relation between the spatial errors computed with our algorithm and the annotations.

4.4. Validation of the temporal lags

Since the expert coaches usually make their evaluation of tennis serves mainly on postures, such as at the key times described in section 4.2, the temporal evaluation was only performed on the karate dataset.

To perform a *tsuki* motion, the right fist is moved forward in a direct path toward the target, with the palm oriented upward at the beginning of the motion and suddenly twisted at the end to finish oriented downward. At the same time, the left fist is moved back to the armed position near the hip. There are 2 main temporal coordinations: the arm displacement synchrony and the simultaneous rotations of both wrists. In this study, we focused only on the arms displacements as the angular rotations of the wrist were not available. An expert's and a novice's motions are illustrated in Figure 11. This example illustrates the correct synchrony of both fists for the expert. On the contrary, the left fist of the novice is only starting to move when the right fist is nearly at the end of its displacement between times t_4 and t_5.

Let $ST_{Ann,i}$ be the annotation of the temporal synchrony between arms l_1 and l_2 during the i^{th} motion and $ST_{Us,i} = \frac{1}{M_0}\sum_{t=1}^{M_0} E^{l_1,l_2}_{Temp,i}(t)$ the global temporal lag we obtain for the same i^{th} motion. $ST^{l}_{Ann,i}$ was based on a scale varying from -10 to 10, the sign reflecting whether the right arm or the left arm is delayed with respect to the other. The higher the lag the higher the amplitude of the annotated score. The comparison between the 14 annotated scores and the estimated temporal lags is illustrated in Figure 12. Once again, a leave-one-subject-out cross validation was used to evaluate the temporal lag of experts' motions.

We can first observe that experts have a perfect temporal synchronization between their arms since their score was around 0 with a small variability. This statement attests that this criterion is fundamental to the right execution of a *tsuki* motion. On the opposite, novices have a large variability with scores ranging from -6 to 10, the sign providing information on which arm is late or in advance (this large variability of the novices has also been observed in the spatial evaluation).

We can also note that the novice motion N_6 does not fit the tendency of other subjects. This can be explained by the fact that this motion is poorly performed as it can be seen on Figure 9 presenting spatial errors. Thus, measuring the coordination between limbs for this example where the limbs made incorrect motion leads to a bad result. Future works will only consider the synchrony of trials with a correct spatial accuracy.

A correlation can be observed between the estimated temporal lags and the annotation scores ($r = -0.8839$ and $p < 0.05$ excluding N_6). This confirms the coherency between the evolution of our temporal lags and the annotations made by experts even if we deal with positive and negative values that are essential to explain the temporal error made by the novice.

5. Conclusion and Discussion

In this paper, we presented a new approach to automatically evaluate the temporal and spatial performance of sport motions independently of the type of sport or the subject's morphology. It is based on a combination of local and global DTW to extract accurate temporal and spatial errors that can be used to improve performance in training sessions.

This approach was validated by comparing our results with annotations of expert coaches made on temporal and spatial features of the motion. Moreover, we made this validation on both

tennis serves and karate motions to emphasize the genericity of our approach.

Several limitations of our method should however be highlighted and be the direction of future works.

First, the morphological normalization is done globally and is not well adapted to consider inter-limb morphological differences between subjects. Even though those differences may be small, this could slightly affects the results. We are currently including a more complex morphological normalization in our method and we will evaluate the importance of these slight differences.

Second, our method does not explicitly consider the cognitive strategies made by the player. For example, during a tennis serve, if the ball is thrown too high, the player must wait for the ball to be at the right height before finishing his gesture. However, our method is dealing with both the temporal and spatial features of the motion. In this example, the alignment path will have a horizontal stage during the wait but the spatial and the temporal errors will not be impacted by this.

As a perspective, we will work on an automatic and unsupervised classification of the experts' motions inside the database depending on their style. We hope that this will reduce variability and thus give even more accurate results and by the way additional information about the style performed.

In this study, we also showed our algorithm can correctly subdivide a motion into its specific phases. A second perspective is to use a subsequence DTW [1, 18] to evaluate a sequence of motions performed by a subject. We could thus provide an evaluation tool that asks a subject to perform motions one after the other and manage to automatically segment and evaluate each of them. Exporting the system to low cost devices, we could enable a subject to perform motions in front of a capture system and watch his performance and progression in real time through an interactive interface.

The present approach opens wide range of use cases. It can indeed be used to automatically compare the novice's motion of any individual sport to the database of experts without adding knowledge or editing/annotating the experts' motions. But it can also be used to compare a novice or injured player along time to evaluate his/her progression. This method could thus be the core of a generic and automatic training system to be used complementary to traditional training sessions.

6. Acknowledgements

This study was partially supported by the funding of ENS Paris-Saclay. Some data used in this project were obtained from both tennis and karate projects carried out in the M2S laboratory. The authors thank Pierre Touzard, Caroline Martin, Anthony Sorel and Anne-Marie Burns for these supplies.

References

[1] Anguera, X. and Ferrarons, M. (2013). Memory efficient subsequence dtw for query-by-example spoken term detection. In *IEEE International Conference on Multimedia and Expo*, pages 1–6.

[2] Barbic, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J. K., and Pollard, N. S. (2004). Segmenting motion capture data into distinct behaviors. *Graphics Interface*, pages 185–194.

[3] Boulbaba, B. A., Su, J., and Anuj, S. (2015). Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14.

[4] Burns, A.-M. (2013). *On the Relevance of Using Virtual Humans for Motor Skills Teaching : a case study on Karate gestures*. PhD thesis, Rennes 2.

[5] Gong, D., Medioni, G., and Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1414–1427.

[6] Heloir, A., Courty, N., Gibet, S., and Multon, F. (2006). Temporal alignment of communicative gesture sequences. *Computer Animation and Virtual Worlds*, 17(3-4):347–357.

[7] Hofstad, E. F., Vapenstad, C., Chmarra, M. K., Lango, T., Kuhry, E., and Marvik, R. (2013). A study of psychomotor skills in minimally invasive surgery: what differentiates expert and nonexpert performance. *Surgical Endoscopy*, 27(3):854–863.

[8] Jiang, Y., Hayashi, I., Hara, M., and Wang, S. (2010). Three-dimensional motion analysis for gesture recognition using singular value decomposition. In *IEEE International Conference on Information and Automation*, pages 805–810. IEEE.

[9] Kahol, K., Tripathi, P., and Panchanathan, S. (2004). Computational analysis of mannerism gestures. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 946–949. IEEE.

[10] Keogh, E. J. and Pazzani, M. J. (2001). Derivative dynamic time warping. In *In First SIAM International Conference on Data Mining*.

[11] Komura, T., Lam, B., Lau, R. W. H., and Leung, H. (2006). e-learning martial arts. In *Advances in Web Based Learning*, volume 4181, pages 239–248. Springer Berlin Heidelberg, Berlin, Heidelberg.

[12] Kulpa, R., Multon, F., and Arnaldi, B. (2005). Morphology-independent representation of motions for interactive human-like animation. *Computer Graphics Forum, Eurographics 2005 special issue*, 24(3):343–352.

[13] Maes, P.-J., Amelynck, D., and Leman, M. (2012). Dance-the-music: an educational platform for the modeling, recognition and audiovisual monitoring of dance steps using spatiotemporal motion templates. *EURASIP Journal on Advances in Signal Processing*, 2012(1):35.

[14] Martin, C. (2013). *Biomechanical analysis of the tennis serve : relationships with performance and upper limb injuries*. PhD thesis, Université Rennes 2.

[15] Martin, C., Bideau, B., Bideau, N., Nicolas, G., Delamarche, P., and Kulpa, R. (2014). Energy flow analysis during the tennis serve comparison between injured and noninjured tennis players. *The American Journal of Sports Medicine*, 41(11):2751–2760.

[16] Martin, C., Kulpa, R., Ropars, M., Delamarche, P., and Bideau, B. (2013). Identification of temporal pathomechanical factors during the tennis serve. *Medicine & Science in Sports & exercise*.

[17] Morel, M., Kulpa, R., Sorel, A., Achard, C., and Dubuisson, S. (2016). Automatic and generic evaluation of spatial and temporal errors in sport motions. In *International Conference on Computer Vision Theory and Application*, pages 1–12.

[18] Müller, M. (2007). *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[19] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2012). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *Journal of Visual Communication and Image Representation*, pages 8–13. IEEE.

[20] Parameswaran, V. and Chellappa, R. (2003). View invariants for human action recognition. In *International Journal of Computer Vision*, volume 66, pages 83–101. Kluwer Academic Publishers.

[21] Pazhoumand-Dar, H., Lam, C.-P., and Masek, M. (2015). Joint movement similarities for robust 3d action recognition using skeletal data. *Journal of Visual Communication and Image Representation*, 30:10–21.

[22] Pham, M. T., Moreau, R., and Boulanger, P. (2010). Three-dimensional gesture comparison using curvature analysis of position and orientation. In *EMBC'10*, pages 6345–6348. IEEE.

[23] Ramakrishnan, A. S. and Neff, M. (2013). Segmentation of hand gestures using motion capture data. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 1249–1250.

[24] Rao, C., Yilmaz, A., and Shah, M. (2002). View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203 – 226.

[25] Raptis, M., Kirovski, D., and Hoppe, H. (2011). Real-time classification of dance gestures from skeleton animation. In *SCA '11 Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 147.

[26] Reiley, C. E., Lin, H. C., Yuh, D. D., and Hager, G. D. (2011). Review of methods for objective surgical skill evaluation. *Surgical Endoscopy*, 25(2):356–366.

[27] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

[28] Sakurai, K., Choi, W., Li, L., and Hachimura, K. (2014). Retrieval of similar behavior data using kinect data. In *14th International Conference on Control, Automation and Systems (ICCAS)*, pages 1368–1372. IEEE.

[29] Sie, M.-S., Cheng, Y.-C., and Chiang, C.-C. (2004). Key motion spotting in continuous motion sequences using motion sensing devices. In *IEEE International Conference on Signal Processing*, pages 326–331. IEEE.

[30] Sorel, A., Kulpa, R., Badier, E., and Multon, F. (2013). Dealing with variability when recognizing user's performance in natural 3d gesture interfaces. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(8):19.

[31] Veeraraghavan, A. and Chowdhury, A. K. R. (2006). The function space of an activity. In *CVPR*, pages 959–968. IEEE Computer Society.

[32] Wang, S. B., Quattoni, A., Morency, L.-P., and Demirdjian, D. (2006). Hidden conditional random fields for gesture recognition. In *CVPR*, pages 1521–1527.

[33] Ward, R. E. (2012). *Biomechanical Perspectives on Classical Ballet Technique and Implications for Teaching Practice*. PhD thesis, University of New South Wales, Sydney, Australia.

[34] Zhong, S. and Ghosh, J. (2002). Hmms and coupled hmms for multi-channel eeg classification. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, pages 1154–1159. IEEE.

[35] Zhou, F. and De la Torre, F. (2015). Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

[36] Zhou, F. and Fernando De la Torre Frade, F. (2009). Canonical time warping for alignment of human behavior. In *Advances in Neural Information Processing Systems Conference (NIPS)*.

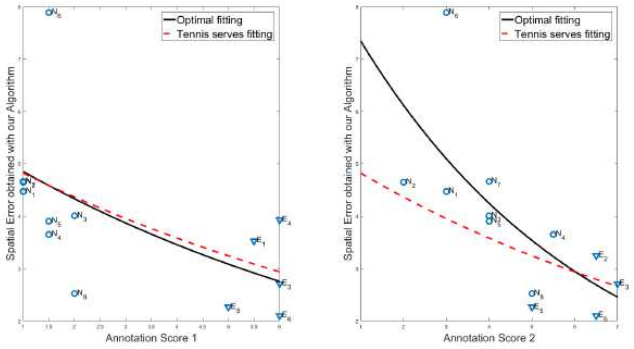


Figure 9: Comparison of spatial error estimation and score annotation of the first coach (left) and the second one (right) for karate *tsuki*. Triangles are for experts and circles for novices. The black line represents the exponential relation between the spatial errors computed with our algorithm and the annotations. The dashed red line is the one obtained with the tennis serves dataset.

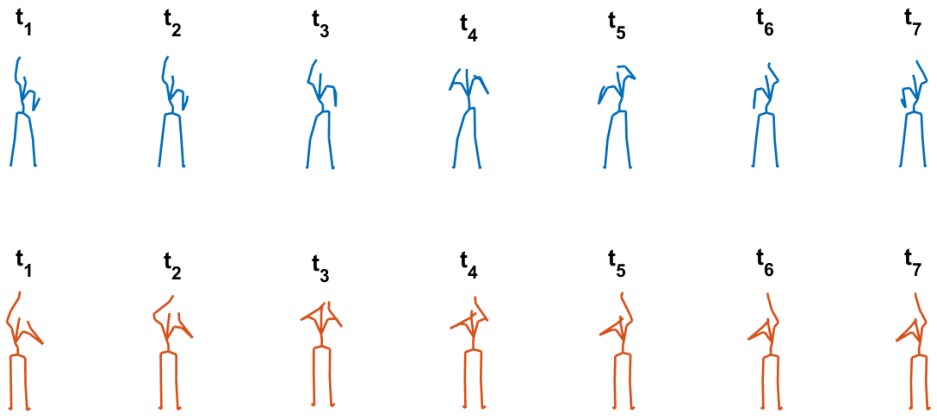


Figure 10: *Tsuki* motions performed by an expert (first row) and the novice N_6 (second row). This Figure highlights the bad positioning of the left arm of N_6 at the end of the motion (see t_7).

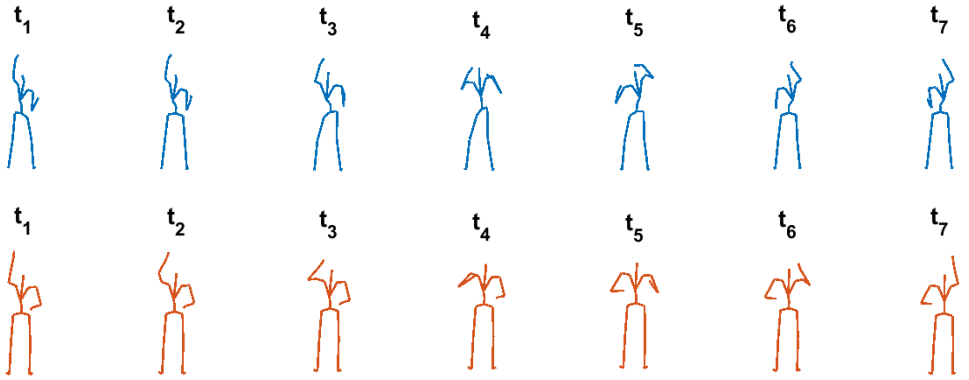


Figure 11: The top (resp. bottom) row shows the *tsuki* motion performed by an expert (resp. novice). This example illustrates the correct synchrony of both fists for the expert. The left fist of the novice is only starting to move when the right fist is nearly at the end of its displacement between times t_4 and t_5.

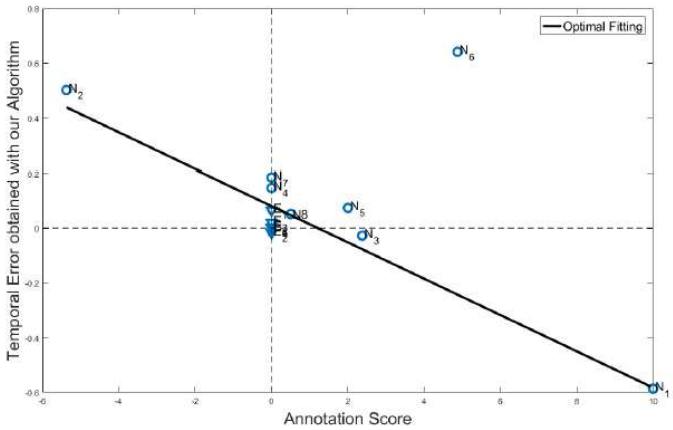


Figure 12: Comparison of temporal lag estimation and score annotation for karate *tsuki*. Triangles are for experts and circles for novices.

Highlights :

- A generic approach of gesture quality estimation based on experts' examples
- An innovative measure of spatial and temporal errors based on a two-level DTW
- Introduction of variabilities in the DTW
- Instantaneous estimation of gesture quality