Karate Moves Recognition from Skeletal Motion

Simone Bianco, Francesco Tisato

DISCo (Dipartimento di Informatica, Sistemistica e Comunicazione)
Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy

ABSTRACT

This work aims at automatically recognizing sequences of complex karate movements and giving a measure of the quality of the movements performed. Since this is a problem which intrinsically needs a 3D model, in this work we propose a solution taking as input sequences of skeletal motions that can derive from both motion capture hardware or consumer-level, off the shelf, depth sensing systems. The proposed system is constituted by four different modules: skeleton representation, pose classification, temporal alignment, and scoring. The proposed system is tested on a set of different punch, kick and defense karate moves executed starting from the simplest case, i.e. fixed static stances (heiko dachi) up to sequences in which the starting stances is different from the ending one. The dataset has been recorded using a single Microsoft Kinect. The dataset includes the recordings of both male and female athletes with different skill levels, ranging from novices to masters.

Keywords: Karate move recognition, skeleton tracking, pose classification, time alignment, Dynamic Time Warping

1. INTRODUCTION

The analysis of complex movements is an important problem for many applications in computer games, sports and medicine,[1] etc. In these applications it is often necessary the description and classification of the skill level in performing different movements and tasks. When practicing a physical skill, it is crucial to have a reassurance that our technique is good or having a feedback on how to correct it. Learning a movement at the level that we can perform it correctly without thinking about it requires lengthy training of "muscle memory", especially if the technique is complex. In sports, a feedback can help athletes in learning what good and bad techniques feels like. As they may not be able to feel whether the whole performance is correct and cannot stop to observe themselves, feedback is required from a third party such as a coach. Furthermore, there are sport competitions in which referees have to judge the athlete's movement giving a score for the technical performance, as for example in diving, skating, gymnastic, etc.

The proposed works aims at automatically recognizing sequences of complex karate movements and giving a measure of the quality of the movements performed. Since this is a problem which intrinsically needs a 3D model, in this work we propose a solution taking as input sequences of skeletal motions that can derive from both motion capture hardware or consumer-level, off the shelf, depth sensing systems. Although it is used in military, entertainment, sports, and medical applications, and for validation of computer vision and robotics, motion capture has some disadvantages. The biggest one is the cost of software, hardware, and equipment needed to acquire the skeletal data. The capture system has to be calibrated for the space it is operated in, and it is not easily portable. On the other hand, consumer level real-time depth sensing systems have known a very rapid diffusion, since the launch of Microsoft Kinect in 2009. Kinect parses a depth-map stream at 30 frames per second to estimate in real-time the positions of 15 predefined points that constitute a wire-frame skeleton of a moving user. Subsequent algorithmic processing can then attempt to understand the user's motion in order to enable him to control and interact through a Natural User Interface (NUI).

The proposed system is constituted by four different modules. The first module is the skeleton representation. This module is used to map the skeleton data to a set of features that enable effective recognition of karate movements. The second module is the pose classification, that models each move as a sequence of key poses. The third module temporally aligns the incoming move with the set of exemplar moves and recognizes which of

Simone Bianco: bianco@disco.unimib.it, Francesco Tisato: tisato@disco.unimib.it

Three-Dimensional Image Processing (3DIP) and Applications 2013, edited by Atilla M. Baskurt, Robert Sitnik, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8650, 86500K

them has been performed. Once the previous module has identified the best matched exemplar move, the fourth module gives a score representative of how well the athlete performed the movement compared to the exemplar one. The system will also be able to give to the athlete a visual feedback on how to improve the movement quality. In the state of the art there are systems fro evaluating the quality of the movement in sports[2,3] and dance,[4] but none of them is able to give a feedback on how to improve the movement quality.

In Section 2 the description of how the skeletal information are extracted is given, together with the explanation of which karate techniques have been considered in this work. In Section 3 the proposed system is described. The experimental results are given in Section 4.

2. SKELETON TRACKING

The Kinect dataset recordings of athletes' performances were captured using the OpenNI[5] drivers/SDK and are OpenNI-encoded (.ONI). The OpenNI SDK provides, among others, a high-level skeleton tracking module, which can be used for detecting the captured user and tracking his/her body joints. More specifically, the OpenNI tracking module produces the positions of 15 joints (Head, Neck, Torso, left and right Shoulder, L/R Elbow, L/R Wrist, L/R Hip, L/R Knee and L/R Foot), together with the corresponding tracking confidence.

The OpenNI tracking module requires a-priori user calibration in order to infer information about the users height and body characteristics. More specifically, skeleton calibration requires the captured user to stay still in a specific calibration pose for a few seconds.

2.1 Techniques

In this work different techniques have been considered. The first ones belong to the category of blocking techniques (see Fig. 1):

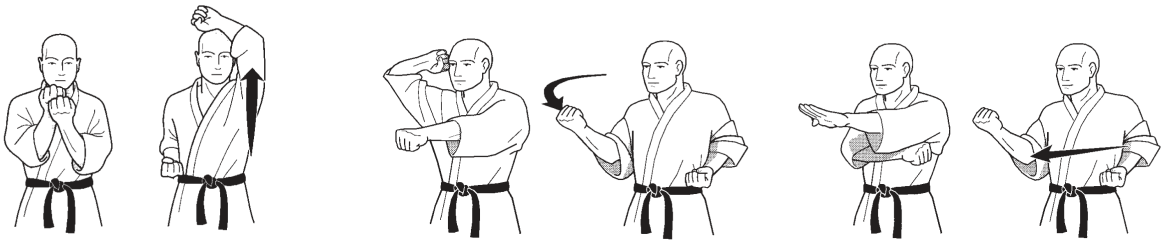
- Age-uke, the rising block.
- Soto-uke, the outside block.
- Uchi-uke, the inside block.
- Shuto-uke, the knife-hand block.
- Gedan-barai, the down block.

The second ones belong to the category of punching techniques (see Fig. 2):

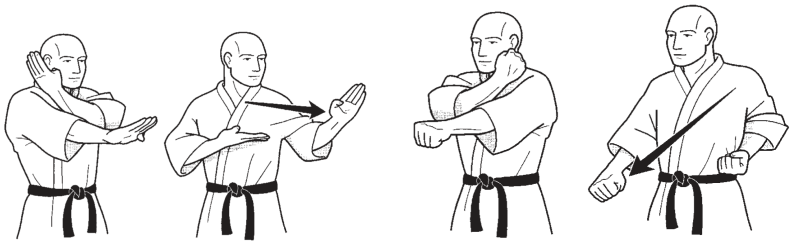
- Choku-zuki chudan, the middle level straight punch (stomach-level).
- Choku-zuki jodan, the upper level straight punch (face-level).

The last ones belong to the category of kicking techniques (see Fig. 3):

- Mae-geri, the front kick.
- Yoko-geri, the side kick.
- Mawashi geri, the round kick.

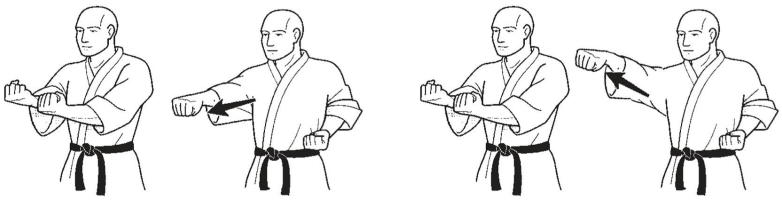


(a) Age-uke (b) Soto-uke (c) Uchi-uke



(d) Shuto-uke (e) Gedan-barai

Figure 1. The blocking techniques chosen



(a) Choku-zuki chudan (b) Choku-zuki jodan

Figure 2. The punching techniques chosen

3. THE PROPOSED SYSTEM

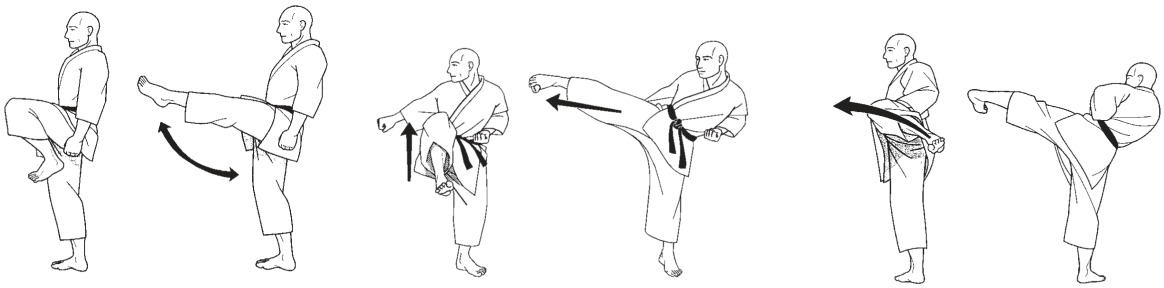
The proposed system is constituted by four different modules: skeleton representation, pose classification, temporal alignment, and scoring. The flow-chart of the proposed system is reported in Fig. 4.

The skeleton representation module is needed to transform the skeleton data acquired by the sensor into a meaningful representation, which has to be invariant from the athlete performing the gesture. The pose classification module classifies each skeleton representation into one of previously defined poses. The temporal alignment module aligns the sequence of vocabulary poses and the template poses in the time domain. The scoring module computes the similarity between two aligned sequences; it can be used both to recognize which move has been performed and to evaluate the effectiveness of the move itself.

3.1 Skeleton Representation

Starting from the acquired skeletal data, we need to recognize which move has been performed and how good it was. We thus need to transform the skeletal data into a normalized representation. This representation has to have the following features:

- be independent from the person acquired;
- be independent from the sensor orientation;
- be compact;



(a) Mae-geri (b) Yoko-geri (c) Mawashi geri

Figure 3. The kicking techniques chosen

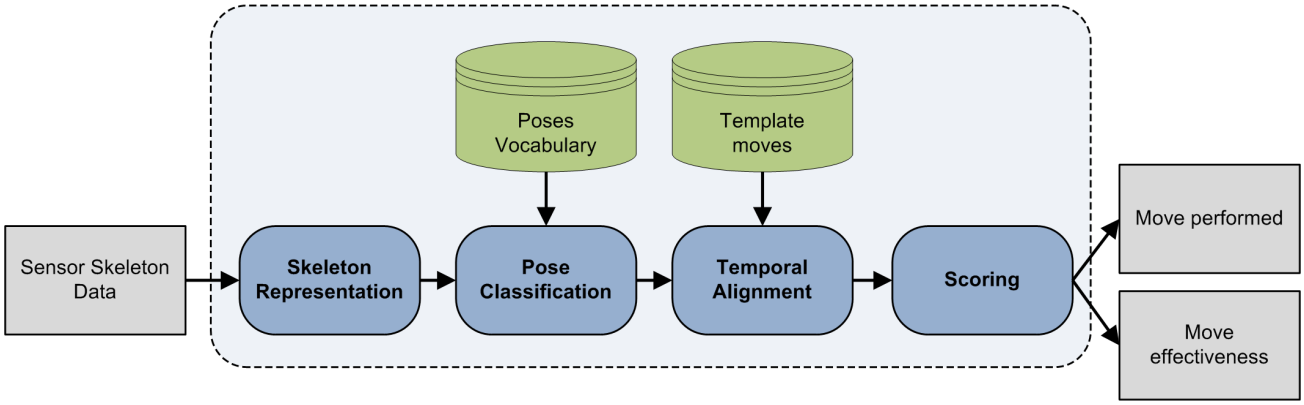


Figure 4. Flow-chart of the proposed system

- maintain useful information in order to be able to give a feedback on the quality of the performed move and how to improve it.

Skeleton coordinates are usually given as 3D coordinates with the origin in the sensor (Kinect data) or with reference to a calibrated volume (MOCAP data). To obtain sensor orientation independence, the first step is to convert the coordinates from the inertial frame of reference to the non-inertial frame of reference centered on the skeleton. The second step is the extraction of descriptive features from the skeleton data: the skeleton joints position have to be described relatively to the corresponding parent joint.

The role of this first module is also to bridge the semantic gap from the sensor model and the model representation of the application domain considered. Thus, we search for a skeleton representation that is close to how the quality of the moves is evaluated. One of the key factors to evaluate the effectiveness of a move is the relative position of the body joints. For example, the position of the elbow is the key to effective blocking. In general, the elbow should finish in line with the hip, or about one fist distance from the ribs. This information can be easily encoded using just one angle, as can be seen in Fig. 5.

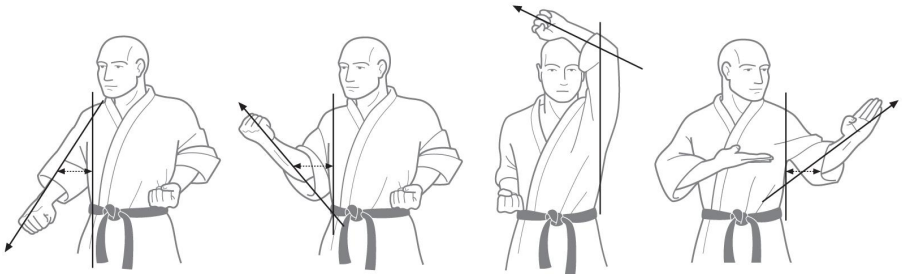


Figure 5. Angles defining the correctness of a blocking technique

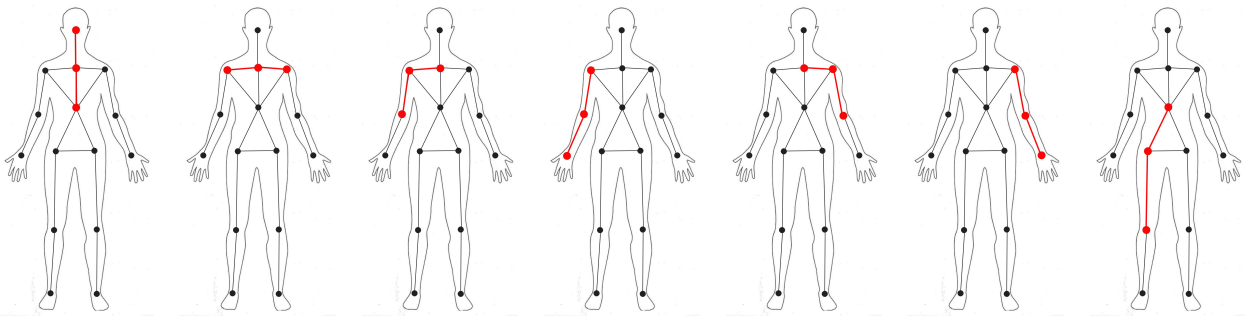
Given the way in which the effectiveness of a move is evaluated, we have chosen an angular representation of the skeletal data. The chosen angles are drawn in Fig. 6. Each angle is defined by three joints J_i, $i = \{s, c, e\}$. Each joint J_i is a 3D point with coordinates $J_i = (x_i, y_i, z_i)$. Before computing the angle defined by the two 3D segments J_sJ_c and J_cJ_e, the three joints are translated so that J_c coincides with the origin O, i.e.

$$\hat{J}_s = J_s - J_c \quad (1)$$
$$\hat{J}_c = J_c - J_c = O \quad (2)$$
$$\hat{J}_e = J_e - J_c \quad (3)$$

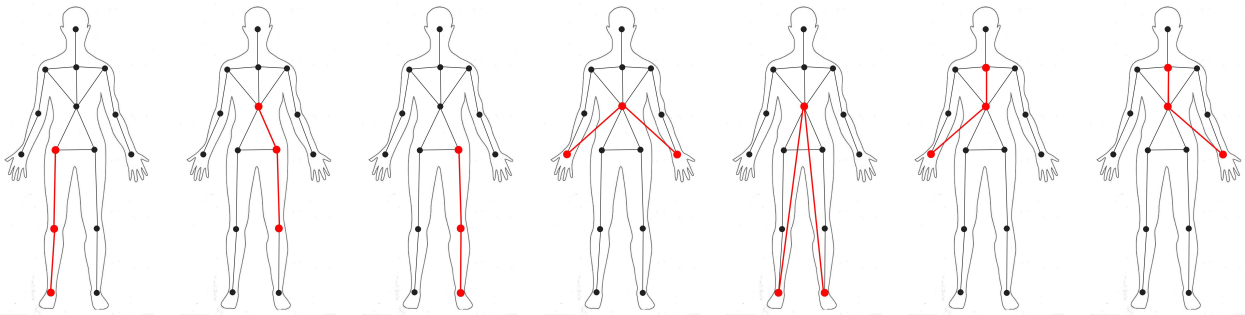
The angle is then computed as

$$\alpha = \cos^{-1}\left(\frac{\hat{J}_s \cdot \hat{J}_e}{||\hat{J}_s||\ ||\hat{J}_e||}\right) \quad (4)$$

The chosen 14 angles are represented in Fig. 6 and are calculated with respect to the central joint.



(a) α_1 (b) α_2 (c) α_3 (d) α_4 (e) α_5 (f) α_6 (g) α_7



(h) α_8 (i) α_9 (j) α_{10} (k) α_{11} (l) α_{12} (m) α_{13} (n) α_{14}

Figure 6. Angles chosen for the skeleton representation

3.2 Pose Classification

The second module is the pose classification. Each move $M = \{M_1, M_2, \cdots, M_s\}$ is modeled as a sequence of key poses p_i, $i = 1, \ldots, s$. Each pose is classified into one of the k predefined key poses $P_1, \ldots, P_k$. The vocabulary of key poses is obtained through a k-means clustering[6] of the moves considered, and each cluster centroid becomes a vocabulary key pose. Two different vocabularies of key poses are obtained for the hand and foot techniques respectively, and then merged together. The classification is done with a multi-class Support Vector Machine (SVM),[7] which recognizes key poses with a one-versus-all approach.

3.3 Temporal Alignment

Given a move, represented as a sequence of poses belonging to a given vocabulary, and a set of template moves represented in the same way, we need to recognize which move has been performed. Since the execution speed can vary from athlete to athlete, we need a module to temporally align the performed and template sequences in order to compute a similarity score that is independent with respect to small to medium velocity changes in the execution.

There are two classes of algorithm in the state of the art that are used for the temporal alignment of sequences: Hidden Markov Models (HMM)[8] and Dynamic Time Warping (DTW).[9] Although there are studies showing that good results can be obtained with both approaches,[8,10–12] in this work we have preferred to use the DTW as it does not require large training sets to learn the template models, but can work even with a single example. In the state of the art exist also correlation-based approaches for the temporal alignment of sequences.[4] The problem with correlation-based approaches is that the sequences can be aligned only by shifting them, without warping. It can be useful for example in dancing applications, where performers have a music to dance on, but is not well suited for our application.

The DTW algorithm permits to calculate the optimal alignment between two sequences. The optimum is obtained minimizing the sum of cumulative distances between the aligned sequences. Formally, given two temporal sequences M and Q with length s and r respectively, i.e.: $M = \{M_1, M_2, \cdots, M_s\}$, $Q = \{Q_1, Q_2, \cdots, Q_r\}$, the DTW distance between the two sequences is found as

$$DTW(M, Q) = \min_{W} \sum_{t=1}^{T} d(w_t) \tag{5}$$

where W is a warping path, i.e. a set of elements

$$w_i = (i_t, j_t) \tag{6}$$

which defines an alignment between the sequences M and Q formally defined as

$$W = \{w_1, w_2, \cdots, w_T\} \quad \text{with } \max(s, r) \leq T \leq s + r - 1 \tag{7}$$

In order for a warping path defined as in Eq. 6 and 7 to represent a feasible solution, it has to satisfy the following criteria:

- Boundary condition: the starting and ending points of the warping path must be the first and the last points of aligned sequences.
- Monotonicity condition: to preserve the time-ordering of points.
- Step size condition: this criteria limits the warping path from long jumps (shifts in time) while aligning sequences.

After defining a dtw matrix of size $s \times r$, where each entry (i, j) contains the local distance $dist(M_i, Q_j)$ between the two corresponding samples M_i and Q_j, it is possible to write the cumulative distance $dtw(i, j)$ recursively as

$$dtw(i, j) = dist(M_i, Q_j) + \min\{dtw(i-1, j), dtw(i, j-1), dtw(i-1, j-1)\} \tag{8}$$

with boundary conditions $dtw(0, 0) = 0$, $dtw(i, 0) = \infty$, $dtw(0, j) = \infty$.

After having computed the optimal DTW distances among the performed sequence and the set of template moves, the performed move is recognized as the one with the lowest DTW distance.

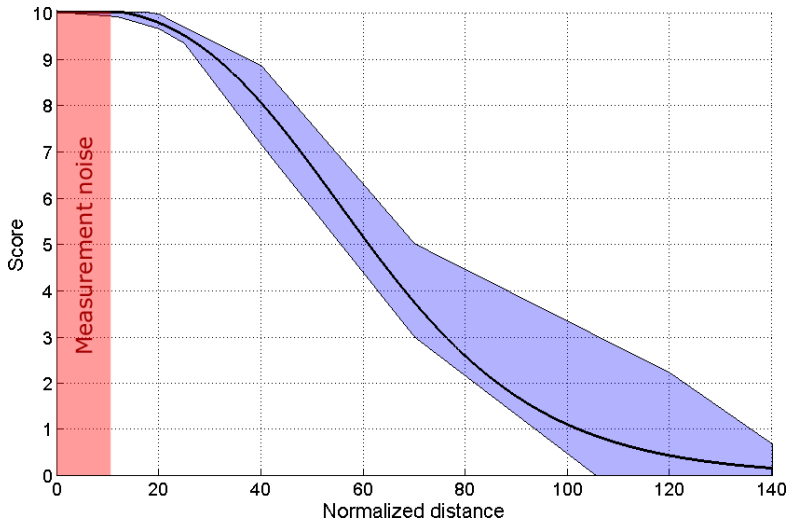


Figure 7. Score computation: the model fitted (black line), median human judgments for each technique executed plus or minus one standard deviation (shaded blue region), measurement noise estimate (shaded red region).

3.4 Scoring

Once the performed sequence has been recognized, we want to give a score of how well it has been performed. The most natural solution could be to use the DTW distance computed in the previous section. The problem with the DTW distance is that it is a cumulative distance, and thus tends to increase for long sequences making it difficult to compare the correctness of sequences of different length. The aim of the scoring module is to give as output a score representative of the effectiveness and quality of the move performed independently of the move length. The score should be in the $[0, 10]$ range, with 10 representing the perfect execution.

We thus define the normalized DTW distance $\bar{d} = DTW/T$, where T is the length of the warping path as defined in Eq. 7. The final score S is obtained by regression among human judgments and the normalized DTW distances $\bar{d}$ obtained by each technique. The model fitted is a 5-parameters logistic:

$$S = \frac{a + fx}{e^{(-bx-c)} + d} \tag{9}$$

In Fig. 7 the model fitted is reported as a black line. The shaded blue region contains the median human judgments for each technique executed plus or minus one standard deviation. The shaded red region fixes a lower bound of the distance $\bar{d}$ obtainable: it is obtained through an acquisition of a completely still athlete and computing $\bar{d}$ with respect to the first acquired frame. It can be thus seen as an estimate of the measurement noise.

The more the first module will be able to bridge the semantic gap between the two different model representations, the more the misalignment score will be informative for the athlete. By analyzing the alignment score along the performed sequence, a feedback can be given to the athlete on how to improve the move execution.

4. EXPERIMENTAL RESULTS

In this work two different experiments have been carried out. In the first one we wanted to test the proposed system is tested on a set of different blocking, punching, and kicking karate moves executed in a fixed static stance (i.e. heiko dachi). The techniques are executed by both male and female athletes with different skill levels, ranging from novices to masters, and include also children. The aim of this first experiment is to assess if the chosen skeleton representation is able to extract the distinctive characteristics of each move, disregarding the physical differences among athletes, such as height, arms and legs length, etc. The athletes were able to freely position themselves in the front of the sensor, at a distance and orientation with the sensor such that the whole body was sensed and to not occlude the technique with their body. The system is tested using as model

	age-u.	soto-u.	uchi-u.	shuto-u.	gedan b.	chudan z.	jodan z.	mae-g.	yoko-g.	mawashi g.
age-uke	1.000	0	0	0	0	0	0	0	0	0
soto-uke	0	1.000	0	0	0	0	0	0	0	0
uchi-uke	0	0	1.000	0	0	0	0	0	0	0
shuto-uke	0	0	0	1.000	0	0	0	0	0	0
gedan barai	0	0	0	0	1.000	0	0	0	0	0
chudan-zuki	0	0	0	0	0	0.925	0.075	0	0	0
jodan-zuki	0	0	0	0	0	0.050	0.950	0	0	0
mae-geri	0	0	0	0	0	0	0	0.975	0	0.025
yoko-geri	0	0	0	0	0	0	0	0	0.900	0.100
mawashi-geri	0	0	0	0	0	0	0	0.050	0	0.950

Table 1. Recognition accuracy for the moves executed in the fixed stance (accuracy=97%)

	age-u.	soto-u.	uchi-u.	shuto-u.	gedan b.	chudan z.	jodan z.	mae-g.	yoko-g.	mawashi g.
age-uke	1.000	0	0	0	0	0	0	0	0	0
soto-uke	0	1.000	0	0	0	0	0	0	0	0
uchi-uke	0	0	1.000	0	0	0	0	0	0	0
shuto-uke	0	0	0	1.000	0	0	0	0	0	0
gedan barai	0	0	0	0	1.000	0	0	0	0	0
chudan-zuki	0	0	0	0	0	0.950	0.050	0	0	0
jodan-zuki	0	0	0	0	0	0.075	0.925	0	0	0
mae-geri	0	0	0	0	0	0	0	0.950	0	0.050
yoko-geri	0	0	0	0	0	0	0	0.025	0.850	0.125
mawashi-geri	0	0	0	0	0	0	0	0.050	0	0.950

Table 2. Recognition accuracy for the moves executed starting from a stance and ending in a different one (accuracy=96.25%)

sequences the recordings from athletes of the World Champion Italian Karate National team. The recognition results are reported in Table 1.

From the results reported in Table 1 it is possible to see that there are some errors. These errors are intra-class for the punching and kicking techniques. A further analysis of the errors reveals that they are relative to acquisitions of novices, and are due to incorrect techniques.

In the second experiment the same set of blocking, punching, and kicking karate moves have been executed starting from stance and ending in a different one. The five stances considered are:

- zenkuthsu dachi, the front stance;
- shiko dachi, the sumo square stance;
- neko ashi dachi, the cat stance;
- sanchin dachi, the hourglass stance;
- moto dachi, the middle-height front stance.

Two different tables are reported for this experiment, in the first one (Table 2) there are the recognition results for the technique executed, in the second one (Table 3) there are the recognition results for the starting and ending stances.

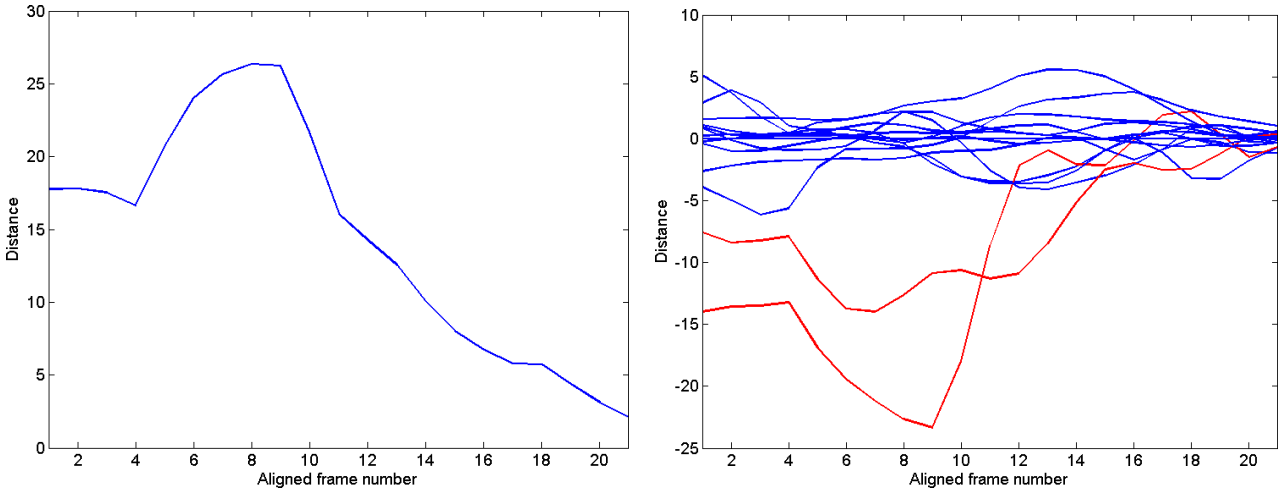
As for the first experiment, the classification errors are intra-class for the punching and kicking techniques. Again, a further analysis of the errors reveals that they are relative to acquisitions of novices, and are due to incorrect techniques.

As last experiment we want to investigate if the proposed system could be used to give a feedback to the athletes to improve their techniques. The case study is the choku-zuki chudan technique in a static stance. The correct technique requires the wrist and the elbow to be aligned towards the punch direction. One of the most common mistakes in the execution of this technique is that the elbow does not follow a straight line,

	zenkuthsu d.	shiko d.	neko ashi d.	sanchin d.	moto d.
zenkuthsu dachi	0.987	0	0	0	0.013
shiko dachi	0	1.000	0	0	0
neko ashi dachi	0	0	1.000	0	0
sanchin dachi	0	0	0	0.950	0.050
moto dachi	0.025	0	0	0.013	0.962

Table 3. Recognition accuracy for the starting and ending stances (accuracy=97.98%)

but executes a curved trajectory. We have selected the acquisition of a novice and compared it with the model sequence. The two sequences have been fist temporally aligned using the DTW algorithm. In Fig. 8(a) the frame-by-frame distance among the temporally aligned sequences is plotted. In Fig. 8(b) the difference for each angle is individually reported. The angle sequence having an absolute maximum difference above the measurement error are identified and plotted in red.



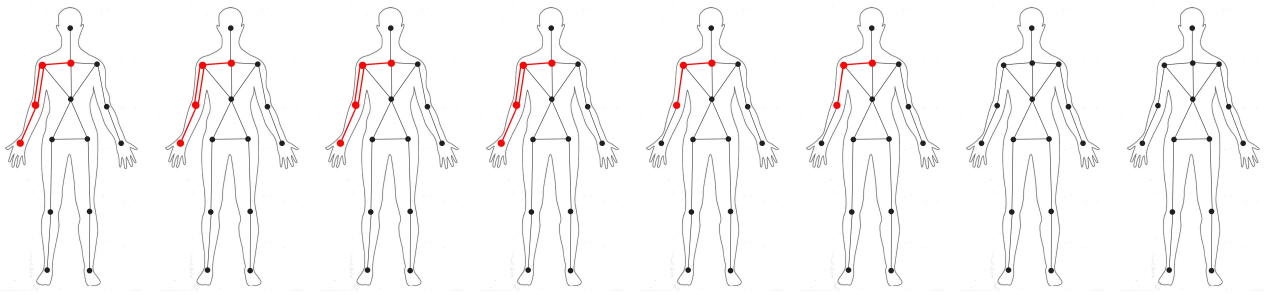
(a) (b)

Figure 8. Frame-by-frame distance among the temporally aligned sequences (a), and individual difference for each angle (b).

The analysis of the aligned distances shows that the part in which the two sequences differ the most is the central one; the initial part is quite similar, and the ending one almost identical. The individual analysis of the aligned differences for each angle (see Fig. 8(b)) shows that the discrepancy is mostly due to two angles, i.e.: α_3 and α_4. The sign of the difference indicates how to perform the correction of the technique: since they are computed subtracting the aligned sequence from the model one, a negative sign means that the corresponding angle has to be narrowed to obtain a more correct technique. A visual feedback can be given to the athlete by playing the acquired move together with a plot in which the angle to be corrected are plotted (see Fig. 9). Two different colors could be used to indicate that the corresponding angle has to be reduced (red, for example) or increased (green, for example).

5. CONCLUSIONS

In this work we have proposed a system which automatically recognizes sequences of complex karate movements and gives a measure of the quality of the movements performed. Our system takes as input sequences of skeletal motions that derive from consumer-level, off the shelf, depth sensing systems. The proposed system is constituted by four different modules: skeleton representation, pose classification, temporal alignment, and scoring. Furthermore, the system is able to give a visual feedback on how to improve the performed techniques. The proposed system is tested on a set of different punch, kick and defense karate moves executed starting from the simplest case, i.e. fixed static stances up to sequences in which the starting stances is different from the



(a) frame 1 (b) frame 4 (c) frame 8 (d) frame 10 (e) frame 12 (f) frame 13 (g) frame 15 (h) frame 20

Figure 9. Visual feedback to the athlete for the correction of the technique: the angles marked in red have to be reduced, while the green ones increased.

ending one. The dataset has been recorded using a single Microsoft Kinect. The dataset includes the recordings of both male and female athletes with different skill levels, ranging from novices to masters.

The system is able to obtain an accuracy of 97% in the case of static stances, and 96.25% in the case of sequences with different starting and ending stances. An analysis of the errors reveals that they are relative to sequences performed by novices, that have a poor movement quality. The system can be used to improve the movement quality during athlete's training, or to teach the movement to a novice.

REFERENCES

[1] Moeslund, "A survey of computer vision-based human motion capture," *Journal of Computer Vision and Image Understanding* **81**, 231 – 268 (2001).

[2] Ilg, W., Mezger, J., and Giese, M., "Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences," in [*Pattern Recognition, 25th DAGM Symposium*], Michaelis, B. and Krell, G., eds., *Lecture Notes in Computer Science* **2781**, 523–531, Springer Berlin Heidelberg (2003).

[3] Chua, P. T., Ventura, D., Crivella, R., Camill, T., Daly, B., Hu, N., Hodgins, J., Schaaf, R., and Pausch, R., "Training for physical tasks in virtual environments: tai chi," *Proceedings of the IEEE Virtual Reality* , 87 – 94 (2003).

[4] Alexiadis, D., Daras, P., O'Connor, N. E., Kelly, P., Boubekeur, T., and Moussa, M. B., "Evaluating a dancer's performance using kinect-based skeleton tracking," *ACM Multimedia* , 659 – 662 (2011).

[5] "Openni." `http://www.openni.org/`.

[6] MacQueen, J. B., "Some methods for classification and analysis of multivariate observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281 – 297 (1967).

[7] Vapnik, V. N., [*The Nature of Statistical Learning Theory*], Springer-Verlag (1995).

[8] Rabiner, L. R., "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE* **77**(2), 257 – 286 (1989).

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

[10] Muller, M. and Roder, T., "Motion templates for automatic classication and retrieval of motion capture data," *SCA* , 137 – 146 (2006).

[11] Muller, M., Baak, A., and Seidel, H.-P., "Efcient and robust annotation of motion capture data," *SCA* , 17 – 26 (2009).

[12] Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A. W., and Campos, M. F. M., "Real-time gesture recognition from depth data through key poses learning and decision forests," *Proceedings of XXV SIBGRAPI Conference on Graphics, Patterns and Images* (2012).