HIGH PERFORMANCE MOVES RECOGNITION AND SEQUENCE SEGMENTATION BASED ON KEY POSES FILTERING

CLÁUDIO MÁRCIO DE SOUZA VICENTE

HIGH PERFORMANCE MOVES RECOGNITION AND SEQUENCE SEGMENTATION BASED ON KEY POSES FILTERING

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Leonardo Barbosa e Oliveira
Coorientador: Erickson Rangel do Nascimento

Belo Horizonte

Fevereiro de 2016

CLÁUDIO MÁRCIO DE SOUZA VICENTE

HIGH PERFORMANCE MOVES RECOGNITION AND SEQUENCE SEGMENTATION BASED ON KEY POSES FILTERING

Dissertation presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Ciência da Computação.

Advisor: Leonardo Barbosa e Oliveira
Co-Advisor: Erickson Rangel do Nascimento

Belo Horizonte
February 2016

Vicente, Cláudio Márcio de Souza

V632h High Performance Moves Recognition and Sequence Segmentation Based on Key Poses Filtering / Cláudio Márcio de Souza Vicente. — Belo Horizonte, 2016
xxii, 41 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Leonardo Barbosa e Oliveira

1. Computação — Teses. 2. Visão Computacional — Teses. I. Orientador. II. Título.

CDU 519.6*82.10(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

High performance moves recognition and sequence segmentation based on key poses filtering

CLÁUDIO MÁRCIO DE SOUZA VICENTE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Leonardo Barbosa e Oliveira - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Erickson Rangel do Nascimento - Coorientador
Departamento de Ciência da Computação - UFMG

Prof. Adriano César Machado Pereira
Departamento de Ciência da Computação - UFMG

Rodrigues Albuquerque
Prof. Maicon Rodrigues Albuquerque
Departamento de Educação Física - UFV

Prof. Thales Miranda de Almeida Vieira
Instituto de Matemática - UFAL

Belo Horizonte, 25 de fevereiro de 2016.

Acknowledgments

First of all, I would like to praise the Lord for He is good and His mercy endures forever. Without Him I can do nothing.

I thank my dear wife Grazielle for being my inspiration and companion during the ours of dedication to the Master's course, my parents for the unconditional support in all matters during all the years of my life and my baby daughter LavÃnia for inspiring me to be a better man.

I also thank my advisor and long term friend Leonardo Barbosa e Oliveira, for encouraging me to apply to the Master's at DCC/UFMG and for all the advices and guidance during the course, my co-advisor Erickson Nascimento for all the Computer Vision knowledge and research experience shared with me.

I am also grateful to all the co-workers that generously helped me with the experiments: Cristiano Gomes Flor, Thales Vieira, all the EEFFTO/UFMG Taekwondo athletes and crew, Luiz C. Emery and Matheus Marques.

In addition I am thankful to Serpro for allowing me to take this course during my working ours, without this support I would not be able to finish the course in an acceptable time.

"For what does it profit a man to gain the whole world and lose his soul?"

(Mark 8:36)

Resumo

Em uma sociedade em que o interesse nos esportes aumenta a cada dia, não é de se surpreender que as pessoas invistam cada vez uma parcela maior dos seus recursos em atividades esportivas. Quando consideramos a prática esportiva profissional, o nível de investimento é consideravelmente maior, já que são detalhes que determinam quem vai estar no topo do pódio. Na busca de melhorar o desempenho dos atletas, uma das áreas que tem atraído o interesse de pesquisadores e na qual os ganhos de desempenho dos atletas têm sido alcançados é a área de análise do movimento humano. Nessa dissertação é apresentada uma metodologia de reconhecimento e segmentação de uma sequência de golpes em esportes de alto desempenho baseado em *key poses* para auxiliar os atletas em seus treinamentos. Quando comparados com gestos humanos realizados no dia a dia, golpes em esporte de alto desempenho são mais rápidos e apresentam baixa variação intra-classe, o que pode produzir um conjunto de características ambíguo e mais propenso a ruídos. Para predizer a classe a qual pertence cada quadro de uma gravação do treinamento de um atleta, nossa abordagem combina uma estratégia robusta de filtragem de quadros compostos por poses discriminantes (*key poses*) com o método probabilístico de reconhecimento de padrões *Latent-Dynamic Conditional Random Field*. A metodologia foi avaliada utilizando sequências de treinamento de Taekwondo não-segmentadas. Os resultados experimentais indicam que nossa metodologia apresenta melhor desempenho quando comparado com o método de reconhecimento de padrões *Decision Forest*, além de ser mais eficiente com relação ao tempo de processamento para reconhecimento dos golpes. Nossa taxa de reconhecimento média foi de 74.72% enquanto o método *Decision Forest* alcançou 58.29%. Os experimentos também mostram que a nossa metodologia é capaz de reconhecer e segmentar golpes realizados em altas velocidades, como os chutes circulares de Taekwondo (*roundhouse kicks*), que podem alcançar velocidades de até 26 m/s.

Palavras-chave: Visão Computacional, Reconhecimento de Ações, *Key poses*, Segmentação.

Abstract

In a society in which the interest in sports increases every day, it is no surprise that people are investing, each time, a grater amount of their resources in sport activities. When we consider professional sport practice, the level of investment is much greater, since the details determine who will be in the top position of the podium. In the search of improving athlete's performance, one of the research areas that have been the focus of attention of researchers and where the athlete's performance improvement have been achieved is the human movement analysis. In this dissertation we present a discriminative key pose-based approach for moves recognition and segmentation of training sequences for high performance sports. Compared with daily human gestures, moves in high performance sports are faster and have low variability inter class, which produce noisy features and ambiguity. Our approach combines a robust filtering strategy to select frames composed of discriminative poses (key poses) and the discriminative Latent-Dynamic Conditional Random Field model to predict a label for each frame from the training sequence. We evaluate our approach on unsegmented sequences of Taekwondo training. Experimental results indicate that as far as processing time and accuracy are concerned our methodology presents the best performance when compared against Decision Forest method. Our average recognition rate was equal to 74.72% while Decision Forest achieves 58.29%. The experiments also show that our approach was able to recognize and segment high speed moves like roundhouse kicks, which can reach peak linear speeds up to 26 m/s.

Keywords: Computer Vision, Action Recognition, Key Poses, Segmentation.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The sport practice has a major importance in modern society as we see people spending time practicing some sport activity for health, beauty and leisure purposes, as well as professionally [Jarvie, 2012]. Due to this importance, we see a lot of institutions making high investments in the sport area (e.g., research laboratories, governments and companies of various fields like clothing, nutrition and medicine industries).

Particularly in Brasil, we have seen the growing interest of the population, media and enterprises due to two of major sport events in which Brasil was selected to host: the Soccer World Cup that happened in 2014 and the upcoming Olympic Games of 2016.

Besides financial investments, the sports development involves the work of professionals of multiple areas, the application of several techniques and a high commitment of the athletes with their training sessions and their physical development [Keegan et al., 2014]. Therefore, high performance athletes are always under a heavy load of training to reach the pinnacle of their physical condition and performance. The heavy load of training they submit themselves, if not properly conducted, can lead to injuries or even career-ending [Carfagno and Hendrix, 2014; Halson, 2014].

Every sport practice involves some kind of movement, and part of the investment made in sports area is being applied in human movement analysis. The biomechanics research area studies and analyses human movement patterns to help the athletes perform some sporting activity better and to reduce the risk of injury, and computer systems have been developed to help biomechanics area analyze the duration, trajectory and strength of the movements performed by athletes.

Human movement analysis is not a new science and the standard method of athlete movement analysis was carried out by video recording, for further human evaluation. This kind of analysis has several problems: 1) it uses hours of video tape recording that require a huge amount of storage. 2) the analysis process is done by manually watching the whole video which is time consuming 3) human analysis is error prone. This process of analysis demands effort that inhibits athletes and coaches to apply it in the daily training routines. Figure 1.1 show an example of this standard method.

Figure 1.1: Example of a standard sport video recording.

With the advances in techniques in Computer Vision and Pattern Recognition fields, however, this task has been made easier. Out of those techniques, two popular ones are: RFID sensors [Wu et al., 2007] and optical markers [Kirk et al., 2005]. Optical markers are required in today's MoCap (motion capture) systems, that are widely used in the movies and games industries, as well as in modern training centers.

The problem with mocap techniques is that they are limited by the necessity of a controlled environment. Usually the mocap application is restricted to a closed environment with controlled lighting. This kind of environment is hard to be implemented and may, in most cases, not reproduce the real condition the athletes face in the usual sport practice. Furthermore, these systems have to be supervised by specialists and employ expensive equipment.

In Figure 1.2[1] we can see the optical markers used in martial arts movement tracking. We can see in this scene that the markers have to be positioned in specific spots of the athletes body and that it is been recorded in a closed and controlled environment.

[1] www.tekgoblin.com/2013/11/13/how-technologies-have-improved-motion-capture

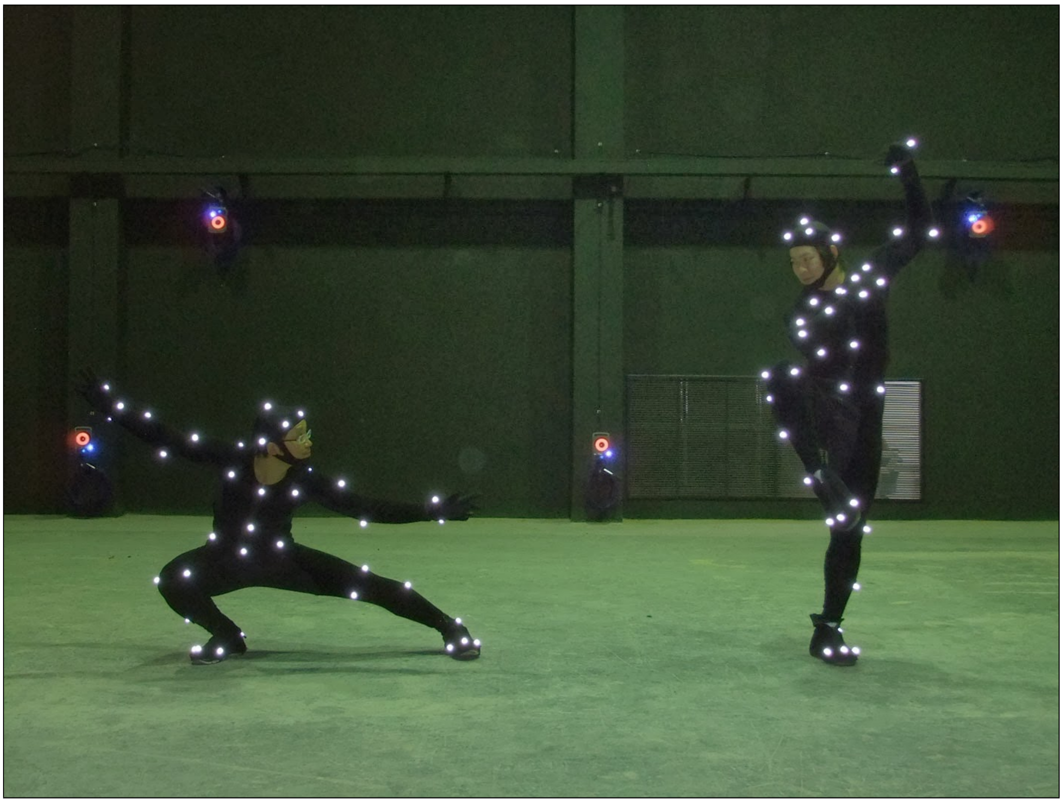


Figure 1.2: Optical markers used in martial arts movement tracking.

In this work, we aim to overcome the drawbacks of the movement analysis techniques previously mentioned and to assist athletes to improve their performance, accordingly. We designed a methodology that does not require a person to analyse the whole video recording to track the moves performed by an athlete in their training routine. Our approach receives a record training sequence and automatically recognizes and segments the moves performed by the athlete. The final segmented video can be used by coaches to analysis the athletic performance in duration and order of each move.

At the same time, our methodology is simple to implement and operate, not demanding a controlled environment neither expensive equipment. To achieve this aim, we employed a low cost RGB-D sensor. In our experiments we specifically employed the Microsoft Kinect. Besides having a low cost, the Kinect has simple operation and can be easily applied in different environments, overcoming the drawbacks of the expensive techniques described previously.

Another important feature of our methodology is that it does not need to store in the hard disk the recording of the whole training routine of the athlete, for further analysis. Instead of using the traditional technique of recording and analyzing all frames of a video, we employed a filtering technique based on key poses [Faugeroux et al., 2014] that selects only a small set of frames to be submitted to analysis. The filtering process is an important part of our methodology since it reduces the amount of information stored and processing time.

Using all frames of a recording to represent fast moves is important when we are dealing with high performance sports. On the other hand it may create a large number of similar frames, which slow down the recognition process and might decrease the classification rate, since the sequence will contain redundant data (a long sequence of frames with repeated or very similar information). This filtering process is actually an adaptive sampling that provides one with a simpler movement representation and consequently a simpler and smaller data to be used in the recognition process.

To eliminate the necessity of a person watch the video for further analysis, we submit the filtered frames of a recording to an action learning method. Specifically, we employ the discriminative Latent-Dynamic Conditional Random Field (LDCRF) learning method [Morency et al., 2007]. The choice of the LDCRF is because it is suitable to recognize sequential and repetitive data, like the moves performed by an athlete during his training routine. LDCRF does it by capturing the intrinsic and extrinsic dynamics of substructure of the moves.

To evaluate our method, we created a new dataset composed of several Taekwondo moves sequences. The dataset was created by recording high performance athletes training sessions at the CTE-UFMG[2] (UFMG's Sports Training Center). The CTE is a modern training center that was designed to enhance the Brazilian general performance in Olympic sports and today it employs physicians, psychologists, nutritionists, supervisors and sports trainers that work with 220 athletes of three different sports: Taekwondo, judo and athletics, with the collaboration of some important high performance training centers around the world.

This approach of combining LDCRF and key poses filtering for fast and high performance moves recognition is a complete novelty, since, to the best of our knowledge, this combination of techniques hasn't been used before in this challenging scenario.

Another contributions of this work are: i) a complete framework for high performance moves recognition and segmentation; ii) to empirically show that using just a small number key frames of a recording may produce better results than using all frames, increasing accuracy and reducing the processing time, and iii) a new dataset composed of high speed moves.

The remainder of this dissertation is organized as follows: in Chapter 2 we present

[2] http://cte.esportes.mg.gov.br/

the works related to action recognition in sports, key poses and LDCRF. Chapter 3 describes the methodology we designed. In Chapter 4 we show the experiments and the results obtained. The conclusion and future works are presented, respectively, in Chapters 5 and 6.

Chapter 2

Related Works

A significant amount of work has been done on gesture recognition and video segmentation. In particular, the video temporal segmentation based on gesture or activities is still a challenge when we consider unsupervised segmentation. The problem of classifying individual frames over time was investigated by Spriggs et al. [2009]. Specific actions from first-person sensor position were analyzed using Gaussian Mixture Models (GMMs), Hidden Markov Models (HMM's), and K-Nearest Neighbor (K-NN). With K-NN on unsupervised segmentation, 61% of the frames were correctly classified. Kernelized Temporal Cuts [Gong et al., 2014, 2012] and Semi-Markov Models [Shi et al., 2008] have also been used in the last years. A modified Hidden Conditional Random Field (HCRF) is used in the work of Wang and Mori [2009] with optical flow features, which require prior tracking and stabilization of the individual in the center of a video sequence. They achieved accuracies from 70.64% to 78.53% in per-frame classification.

These techniques are suitable for gesture recognition and video segmentation, working well on sequential data, but the LDCRF proved to retrieve best results on human motion when compared to them [Morency et al., 2007]. When compared to the HCRF, the advantage of the LDCRF is that it is an evolution of the HCRF. HCRF is trained on sets of pre-segmented sequences and do not capture the dynamics between gesture labels, only the internal structure. The LDCRF, on the contrary, captures both extrinsic dynamics and intrinsic structure.

In the work of Xia et al. [2012] it is presented a method to classify actions using an HMM based solution that relies on posture visual words, built from Histograms of 3D joints (HOJ3D). Devanne et al. [2013] used spatio-temporal motion trajectories to represent human actions and a K-NN classifier was adopted to recognize gestures, achieving better results when compared to the method of Xia et al. [2012]. A depth data based method is proposed by Yu et al. [2014]. The authors present an online approach based on orderlets using both skeleton and depth data to classify frames.

The Hidden Markov Model based recognition techniques mentioned previously are generative models and give their output directly by modeling the transition matrix based on the training data. Their purpose is to maximize the joint probability of paired observation and label sequences. The advantage of choosing the LDCRF is because the CRF based models are considered generalizations of the HMMs, avoiding the label bias problem.

Despite the wide range of work for gesture recognition using low cost devices (e.g., cameras, RGB-D sensors, etc.), virtually all works focus on every day human actions [Elgendi et al., 2012; Vieira et al., 2013; LaViola, 2013; Gavrila, 1999]. There is a small number of works that is tackling with movements in general sports [Gray et al., 2014; Choppin and Wheat, 2013] and even smaller for high performance sports such as martial arts [Bianco and Tisato, 2013]. Wada et al. [2013] developed a motion analysis system designed to help martial arts choreographed movements training.

Similar to these works we will use a RGB-D sensor for human movement recognition, but in a more challenging environment. The works developed in the previous paragraph work on every day human actions. These actions are e.g, lifting an object, waving the hand, nodding. These actions do not have the complexity and fast speed of fight moves like the ones we are dealing in this work. These fast moves require different learning and recognition techniques of simpler movements. For this reason we developed a completely new recognition and segmentation methodology.

The small number of works that dealt with sport movements is also a stimulus for our work. Usually these works studied the sport movements in a softer way that is, a person executes a movement in a certain way that an athlete would not do in his daily activity. Even when the works approached martial arts, they did not applied it in the way the athlete would perform it in a competition.

In general, moves recognition in sports context [Bialkowski et al., 2013; Gade and Moeslund, 2013; Pers and Kovacic, 2000; Hu et al., 2014; Hamid et al., 2014; El-Sallam et al., 2013] has been tackled with complex setup of cameras. El-Sallam et al. [2013] presented a 24 opto-sensitive cameras setup capturing 50 frames and a markerless system to be used to optimize the athletes techniques during training sessions. Their system aims sports such as jumping, throwing, pole vault, and javelin throw. The methodology can also be applied in other sports that do not essentially require the tracking of the full joint kinematics, since it is focused on velocity evaluation and silhouette segmentation. In addition to the 24 cameras, it requires a manual identification of the movements.

The 24 opto-sensitive cameras setup used in El-Sallam et al. [2013]'s work is the setup we intend to avoid in our work. This setup has the drawbacks cited previous that we want to overcome. It requires expensive equipment and controlled environment which not only restricts the system to few high income training centers but also uses an environment different from the one the athlete experience in the sport practice. In our methodology we apply a single inexpensive RGB-D sensor which does not require a controlled environment.

Another difference from El-Sallam et al. [2013]'s work is concerning the identification of the movements. El-Sallam performs a manual identification of the movements while our focus is in automatic action recognition and segmentation. The similarity with our work is that we also use a markerless approach in sports context.

For daily gestures or moves in sports, a common procedure to compute features is to analyze the human pose. Several approaches have been proposed to estimate human poses. A non-rigid articulated point set registration framework for human pose estimation was developed by Ge and Fan [2015]. The authors developed the technique to improve two registration techniques, Coherent Point Drift (CPD) and Articulated Iterative Closet Point (AICP). Zecha and Lienhart [2015] worked on a new methodology for detecting key poses in top-class swimmers videos. They used RGB side images of swimmers to manually annotate the key poses for later recognition using a maximum likelihood approach which predicts the key-pose.

We will also use key poses in a sport context but, differently from Zecha and Lienhart [2015]. In their work, the key poses are manually annotated. They have to record swimmers videos, chose which of the several poses of a swimmer which ones are the key ones and submit them to the recognition method. In our methodology the key poses are detected without any human intervention. The automatic key poses extraction has advantages over human extracted key poses in the way that it always gives the same output if the same input is given. Human chosen key poses are not always the same given the same input, since each person may consider their own subjective factors.

Zecha and Lienhart [2015] also worked on key poses for only one type of movement while we work with key poses of several different moves. One of the difficulties of key poses extraction arise when we have to extract key poses of several movements. Many movements start and finish at the same pose, but have different poses in between. Our work involves the difficulty of extracting key poses of moves that only differ in the very ending segment while Zecha and Lienhart [2015] did not have to deal with this issue.

Although the Kinect sensor we used in our methodology does provide RGB images, we do not use this kind of information in our recognition method. The Kinect provides us with a skeleton tracking that gives an estimation of the athlete skeleton in the 3D space. The fact that Zecha and Lienhart [2015] had to use the RGB image comes from the presence of water in the scene, that prevents Kinect from estimating the swimmer skeleton.

Borras and Asfour [2015] proposed a taxonomy of whole-body balancing poses using the environment to enhance the poses identification. They covered a wide variety of poses specifically designed for humanoid robots. The most significant difference from our work is that they do not include any kind of fighting poses. Their poses are related to slow movements of humanoid robots. Another difference from our work and Borras and Asfour [2015]'s work is that we don't consider the environment to make the poses extraction. In future works we can actually consider the environment for moves recognition. In many training sessions, the Taekwondo athelet uses apparels like rackets. This kind of environment object can be useful in key poses extraction and moves recognition.

Another robotic application is the one designed by Lea et al. [2015]. They used a Skip-Chain Conditional Random Field to automatically segment and recognize fine-grained robot surgical in a constrained environment. A Skip-Chain CRF is a pure CRF learning method with an additional feature of skipping some elements of sequential data. They initially submit all the data to the CRF method and further process the elimination of some elements of the CRF chain of sequential data. Our methodology works differently in the fact that the filtering step occurs before submitting the data to the learning method. Our prior filtering is more efficient than eliminating elements of the CRF chain after submitting all the data to the recognition method.

Similarly to our work, Faugeroux et al. [2014] also use a subset of frames extracted from the sequence. In their technique, a decision forest is built by using nodes to represent discriminative key poses and leaves to represent gestures, i.e. moves in our case. Each path in a tree of the forest, from the parent of a leaf to the root, corresponds to a possible sequence for a move in reverse order, to allow online recognition. Therefore, each tree rooted at a certain key pose k encodes all moves whose final key pose is k. In this way, the maximum number of trees is the number of key poses in the key pose set. Due to this similarity we will use this work as our baseline in the experiments.

Chapter 3

Methodology

The input of our method is a training set composed of sequences of frames $\{f_1, \ldots, f_n\}$ and their respective move labels represented by sequences of moves labels $\{y_1, \ldots, y_n\}$. Each label is a member of a set $\mathcal{Y}$ of possible moves labels, e.g. $\mathcal{Y} = \{FrontLegKick, BackArmPunch \ldots\}$. Each frame f_i is represented by a feature vector $\Phi(f_i) \in \mathbb{R}^d$, where the function Φ extracts features by analyzing the human pose.

Our first objective is to perform a supervised learning using this training set, by training a LDCRF model. Using this learning approach, we aim at recognizing and segmenting unlabeled sequences containing several executions of different moves.

Our training methodology can be briefly described in the following steps:

1. key pose extraction from the training set data;

2. key pose labeling and filtering of each trained sequence;

3. LDCRF training using the filtered frames represented by their feature vectors computed by Φ.

The model generated by the LDCRF will be used for moves recognition, as described in Section 3.3. Figure 3.1 illustrates the whole process.

3.1 Key poses extraction

The first step of our method is the selection of the most discriminative frames (key poses) from the sequence. In the work of Faugeroux et al. [2014] is presented a methodology to automatically extract a concise and discriminative key pose set as well as the key pose sequence representing each gesture example. To extract the key poses, the

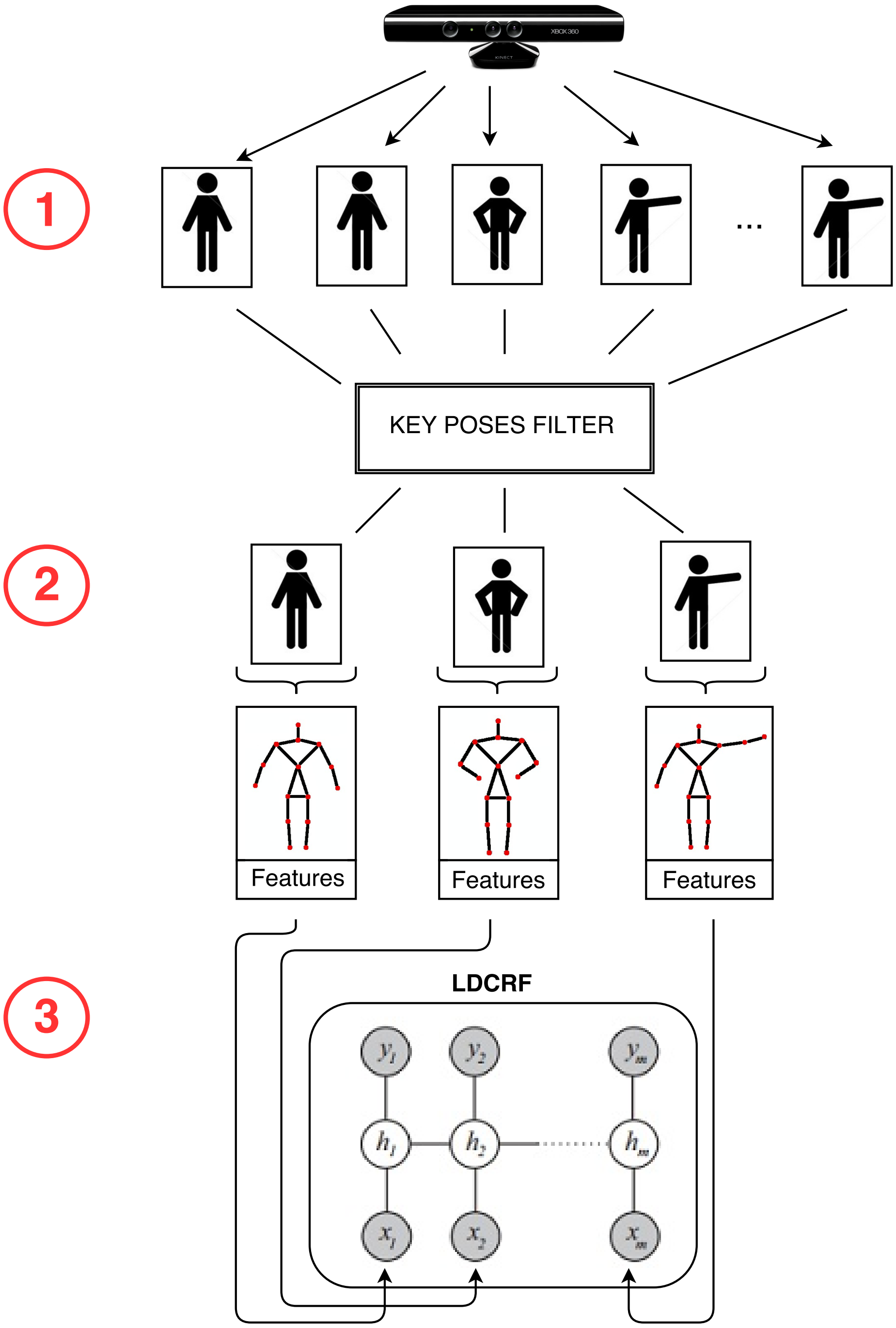


Figure 3.1: Graphical representation of our methodology: 1) Skeleton data is extracted by a RGB-D sensor; 2) These frames are filtered and a set of frames containing key poses are selected; 3) Training and testing with the acquired features using the discriminative LDCRF model.

method initially represents a record of gestures (several gestures performed with a short pause between each one) as a curve in high dimensional space, and consider that between each gesture execution, the user performs short pauses. These pauses intervals can be easily identified by intervals of high curvature and the gestures are characterized by low curvature intervals, allowing a robust segmentation of gestures. All the process is carried out using a skeleton estimated from depth.

After the moves segmentation, the key poses can be extracted using a greedy strategy as follows:

1. The first and last pose of each gesture are marked as key poses. If a moves starts and ends at the same pose, the pose that differs most from both the initial and final poses is also added;
2. If moves with the same representation are found, then discriminant poses are added to those moves‘ sets until all representations become unique. The Hausdorff Distance between gestures is used to select the most discriminative key poses.

Given two finite point sets $P = \{p_1, \ldots, p_p\}$ and $Q = \{q_1, \ldots, q_q\}$, the Hausdorff Distance is defined as [Huttenlocher et al., 1993]

$$H(P, Q) = max(w(P, Q), w(Q, P)), \tag{3.1}$$

where

$$w(P, Q) = \underset{p\,\in\,P,\;q\,\in\,Q}{max\;min} ||p - q||, \tag{3.2}$$

and $||\ .\ ||$ is some norm (e.g., L_2 or Euclidean norm) on the points of P and Q.

The key poses must be trained a priori, i.e. we must first define the group of moves we want to analyze and thus submit them to the key poses extraction process. Each key pose from frame f_j is represented internally by a feature vector $\Phi(f_j)$ that is the concatenation of two types of skeletal information, namely:

- **Joints:** the 3D coordinates of each of the 15 body joints;
- **Joint-angles:** 9 zenith θ and 8 azimuth ϕ angles of upper and lower body joints [Miranda et al., 2012].

The final feature vector is given by

$$\Phi(f_j) = (J_1, \ldots, J_{15}, \theta_1, \ldots, \theta_9, \phi_1, \ldots, \phi_8)^T, \tag{3.3}$$

where $J_i \in \mathbb{R}^3$ is the i-th joint, θ is the zenith and ϕ the azimuth of the skeleton extracted from key pose of frame f_j. The elements of the feature vector used in the learning methods are described in the Chapter 4.

3.2 Filtering process

The filtering step is an important part of our methodology. Traditional video recording have a frame rate of around 30 frames per second. Although there are camera systems that work in a very higher frame rate a 30 FPS rate already produces a huge amount of information to be stored and submitted to learning methods. Even in high speed moves like the ones stroke by Taekwondo athletes, we notice that several contiguous frames of a video recording have body poses with very little variation among them. Due to these very similar contiguous frames, we initially work on the hypotheses that we could harm the model generated by a learning method that maps sequential data, like the LDCRF, if we submit all of these frames to it.

Our filtering process in on-line, that is, we analyse the frames that are being submitted by the RGB-D sensor and stores only the ones that satisfy our filtering algorithm. This procedure effectively reduces the storage space necessary to learn and recognize a move that is being performed by the athlete since we do not need to store all frames that are being processed by the system in the computer hard disk for further filtering. Although this process is on-line, it can also be done with a unfiltered recording that was previously stored in the hard disk.

This process starts after the key pose extraction. Once the key pose extraction is completed, we have a set $\mathcal{K} = \{k_1, k_2, \ldots, k_m\}$ of key poses, that group all key poses previously extracted from all moves. We take the complete sequence of frames $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ and map each one to a key pose of our set of key poses $\mathcal{K}$ using a nearest-neighbor classifier combined with the pose similarity threshold $\epsilon = \pi$ to classify a frame f_i. If there is no key pose assigned to a frame, it returns an invalid pose. The k Nearest-Neighbor (k-NN) classifier is given by the function:

$$g(f) = \begin{cases} k_p = \underset{k \in \mathcal{K}}{\text{argmin}} \Delta(k, f) & \text{if } \Delta(k_p, f) < \epsilon, \\ \text{-1} & \text{otherwise,} \end{cases} \tag{3.4}$$

where Δ is the distance function between the key pose k and the pose in the frame. The Δ function used in our experiments is the geodesic distance on the sphere, that is used to compare the same joints in two distinct poses, using their 9 spherical angles:

$$\Delta(k, k') = \sum_{a=1}^{9} [\delta(k_a, k'_a)]^2, \tag{3.5}$$

where

$$\delta(k_a, k'_a) = arcos(sin\ \theta_t\ sin\ \theta'_t + cos\ \theta_t\ cos\ \theta'_t\ cos\ |\phi_t\ - \phi'_t|). \tag{3.6}$$

Considering the previously mentioned information that an athlete performs a move at 30 FPS, there will be little body pose variation among a sequence of contiguous frames, even for fast moves. Therefore, it is easy to notice that our methodology will assign the same key pose to a sequence of contiguous frames until another the pose of a certain frame becomes more similar to another key pose. So, we can properly select only one frame, out of the contiguous frame sequence that has the same label, to be the representative frame of that interval.

At the end of this process, we will have our filtered frames set $\mathcal{X} = \{f_1, f_2, \ldots, f_i\}$, in which we can associate a move label to each frame, forming our moves set $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$.

Although simple, this frame selection strategy proved to be efficient, as will be shown in Chapter 4. From the sequence of frames that have the same key pose label, we choose the first one in its sequence. The choice of the first one of the sequence was arbitrary and make according of our feeling of what frame of the sequence would return good results. Tests using other frames, like the middle one, are described in the Future Works chapter. In this way, instead of saving the skeletal information at each frame, we have a considerably smaller amount of information saved at each recording, which will be further used in the learning process.

3.3 Model training

One benefit of the filtering process is to have a less amount of information to store, reducing the processing time. Besides, one of our goals is to evaluate the performance of the recognition using a smaller number of frames compared to the all frames approach. Thus, for each move performed by the athletes we extracted 17 angles (θ and ϕ) of upper and lower body joints (according to [Miranda et al., 2012, 2014]) per frame at a 30 FPS rate.

The recognition is performed by a LDCRF model trained with training data. LDCRF is a discriminative method that combines the CRFs capabilities of capturing

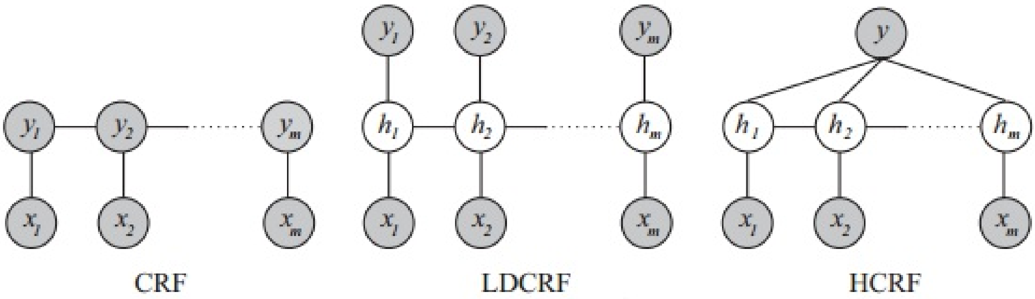


Figure 3.2: Comparison of the LDCRF model with the CRF and HCRF models. With x_n as observations, h_n the hidden variables and y_n the labels

extrinsic gesture dynamics with the HCRFs capabilities of capturing intrinsic gesture sub-structure. It generates hidden state variables that model both the sub-structure of a class sequence and the dynamics between class labels [Morency et al., 2007].

A CRF is a type of discriminative probabilistic model that maps the relationship between observations and is often used for labeling sequential data. The HCRF is an important evolution of the traditional CRF since it estimates a class given a segmented sequence. For this purposes it generates a classification model with hidden variables (states). The LDCRF also works with hidden variables like the HCRF, but instead it does not need a segmented sequence to map a class, it is able to map a class to each observation and generating a hidden variable to each one.

LDCRF generates an undirected graphical model whose nodes can be divided in three disjoint sets: i) the observed variables; ii) the hidden variables (latent variables) and iii) the output variables, labels. Figure 3.2.

While an ordinary classifier predicts the label of a sample regardless its neighbor samples, CRF-type classifiers (CRF, HCRF and LDCRF) will take into account the relationship between neighbor samples, that is, it takes the context into to make its label predictions.

Contrary to generative models, that assume an independence among observation, LDCRF assumes that observations have a dependence among them being more adequate to map long-range dependencies among observations. This kind of discriminative model suits well with supervised learning, for this reason, in our training phase we explicitly define our inputs and their corresponding outputs.

Another advantage of LDCRF is its capabilities of making a per frame classification. By using LDCRF we are capable of i) mapping between a sequence of frames and their respective labels and ii) an association of disjoint sets of hidden states with each class label. Therefore, LDCRF will enable a mapping between the filtered frames of a recording and their given Taekwondo moves.

After filtering the frames as described in Section 3.2, we submit our set of filtered frames $\mathcal{X}$ and the corresponding set of labels $\mathcal{Y}$ to the LDCRF. The learning step generate the model, creating a vector of hidden variables $\mathbf{H} = \{h_1, h_2, \ldots, h_m\}$ for each pair key pose/label. This vector estimates the sub-structure of the moves sequence. The final latent conditional model is given by

$$P(\mathbf{y}|\ \mathbf{k},\ \Theta) = \sum_{h\ \in\ H} P(\mathbf{y}|\,\mathbf{h},\ \mathbf{k},\ \Theta)\ \ P(\mathbf{h}\,|\,\mathbf{k},\ \Theta), \tag{3.7}$$

where Θ is the vector with the parameters model, $\mathbf{h}$ is a member of a set $\mathbf{H}$ of possible hidden states for each pair key pose/label and $\mathbf{k}$ is the set of key poses represented by feature vector, i.e. $\mathbf{k} = \{\Phi(k_1), \ldots, \Phi(k_m)\}$.

Since the LDCRF has $P(\mathbf{h}\,|\,\mathbf{k}, \Theta) = 0$ when the sets of hidden states H are disjoint, we have:

$$P(\mathbf{y}|\ \mathbf{k},\ \Theta) = \sum_{h\ \in\ H} P(\mathbf{y}|\,\mathbf{h},\ \mathbf{k},\ \Theta). \tag{3.8}$$

Using the CRF formulation we can write $P(\mathbf{y}\,|\,\mathbf{h},\ \mathbf{k}, \Theta)$ as:

$$P(\mathbf{y}|\ \mathbf{k},\ \Theta) = \frac{1}{Z(\mathbf{k}, \Theta)} \exp\left(\sum_{j=1}^{m} \Theta_j \cdot F_j(\mathbf{h}, \mathbf{k})\right), \tag{3.9}$$

where $Z(\mathbf{k},\ \Theta)$ is the normalization function:

$$Z(\mathbf{k}, \Theta) = \sum_{h\ \in\ H} \exp\left(\sum_{j=1}^{m} \Theta_j \cdot F_j(\mathbf{h}, \mathbf{k})\right), \tag{3.10}$$

and the potential function $F_j(\mathbf{h},\ \mathbf{k})$ is either the sum of the transition functions or the sum of the state functions.

During test phase, whenever we submit a new set of filtered key poses, the LDRCF estimates the most likely label y^* that maximizes the conditional model:

$$y^* = \arg\max_{y} \sum_{h\ \in\ H} P(\mathbf{h}\,|\,\mathbf{k},\ \Theta^*), \tag{3.11}$$

where Θ^* represents the parameters of the trained model.

Chapter 4

Experiments

We evaluated the performance of our approach using a dataset composed of Taekwondo moves. Taekwondo was the chosen sport due to its emphasis on speed and agility, features we wanted to investigate and that make our work different from the ones described in the related works. Taekwondo has also a great variety of moves that would be suitable to evaluate the correctness of key poses extraction. Another factor that contributed in our choice of Taekwondo as the chosen sport is the CTE's availability in providing us with high performance athletes to experiment our methodology.

All data were acquired using a single Microsoft Kinect (a low cost RGB-D sensor). The Kinect is a device that can provide to its user the following resources [Catuhe, 2012]: i) Depth map image with color gradients, ii) RGB 30 FPS images, iii) Multiple skeleton tracking (at most 6 persons) in the same scene, iv) 20 joints 3D coordinates according to the Figure and v) Precise microphone array. Among all the resources it provides we will use the combination of skeleton tracking and 15 joint 3D coordinates, which will allow us to track the athlete's body and build the feature vector (Section 3.1) of each frame of the recording.

We selected seven Taekwondo moves in our experiments, namely: i) front leg kick, ii) back leg kick, iii) front leg kick to the head, iv) back leg kick to the head, v) front arm punch, vi) back arm punch, and finally vii) spinning kick. These seven moves were the moves recommended by the CTE's Taekwondo head coach to our experiments. These moves are commonly used by the athletes in their daily training sessions. In addition, this moves set cover some important characteristics we would like to investigate in our research since it includes completely different execution moves (spinning kick and back arm punch) and moves which execution only differs in the vary final segment (front leg kick to the headand front leg kick).

A characteristic that highlights our approach from other recognition and segmentation works is that we perform experiments with high speed movements like roundhouse kicks [Mailapalli et al., 2015], which can reach peak linear speeds up to 26 m/s.

4.1 Data acquisition

Our experiments were carried out at the sports training center with the guidance of its Taekwondo head coach. We selected ten athletes of different sex, ages and heights to perform the seven different moves. The athletes ages ranged from 14- to 26-years old and their height from 160 cm to 176 cm. We can see the detailed athletes' body characteristics in Table 4.1.

Athlete	Sex	Age	Height (m)	Weight (Kg)	Belt Color	Experience Level
Athlete 1	Male	14	1.60	45.15	Blue	Beginner
Athlete 2	Female	17	1.61	53.25	Red	Advanced
Athlete 3	Female	22	1.65	62.40	Black	Advanced
Athlete 4	Male	18	1.67	64.40	Red	Intermediate
Athlete 5	Female	16	1.71	52.10	Blue	Intermediate
Athlete 6	Male	20	1.72	66.15	Black	Advanced
Athlete 7	Female	18	1.73	74.25	Red	Intermediate
Athlete 8	Male	21	1.74	74.60	Red	Advanced
Athlete 9	Male	26	1.74	78.75	Black	Advanced
Athlete 10	Male	17	1.76	72.85	Yellow	Beginner

Table 4.1: Taekwondo athletes' characteristics.

During the data acquisition, the RGB-D sensor was set to a position where its camera and sensors could capture the athlete's body sideways. The Figure 4.1 shows how the Kinect and the athlete body were positioned during the moves capture while Figure 4.2 shows the sensor perspective of an athlete during a move execution.

Since there is little variation of each move among athletes (they are professional athletes), the extraction of key poses was performed using the training sequence from a selected single athlete. Figure 4.3 shows the seven moves and their respective extracted key poses. The remaining sequences are used for the recognition and segmentation processes.

We evaluated the recognition rate of the two kinds of features, namely: All-frames (unfiltered) and Key-poses. Beyond comparing the filtered and unfiltered recognition rate, we contrast our methodology and Decision Forest.

The Decision Forest is a learning method that uses multiple decision trees to build a classifier [k. Ho, 1995]. It uses the simple training and high classification speed

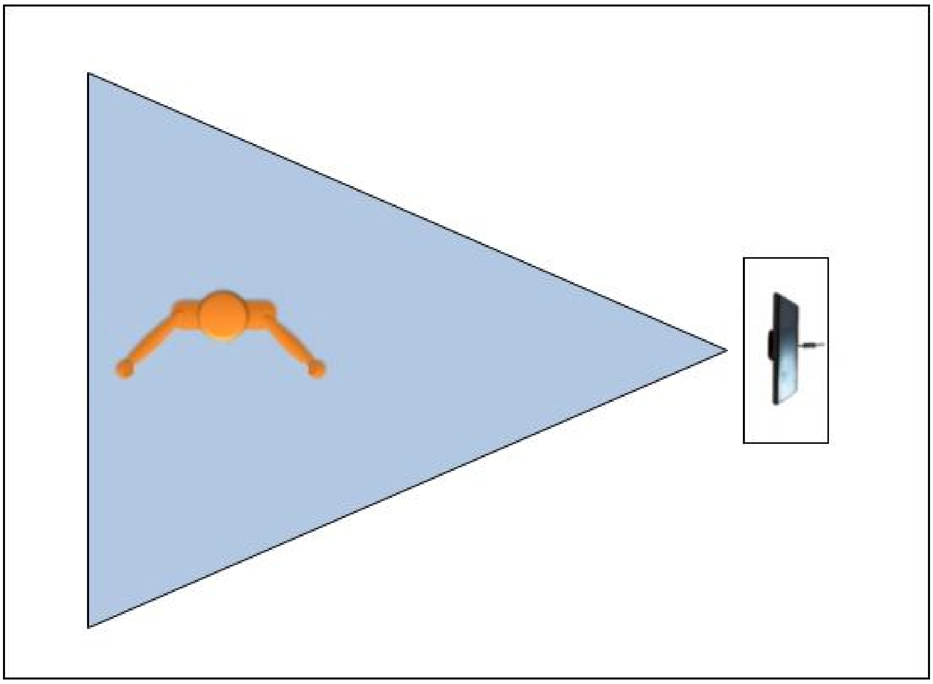


Figure 4.1: Moves recording setup.

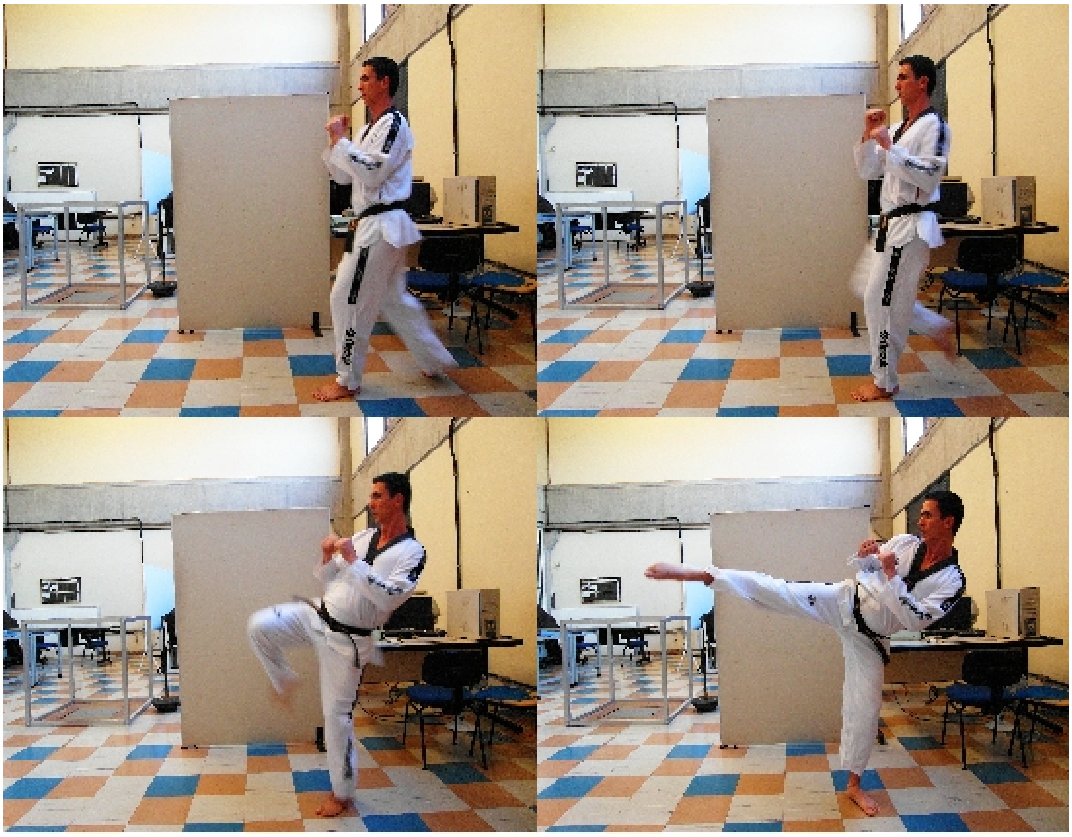


Figure 4.2: Sensor perspective of an athlete during a move execution.

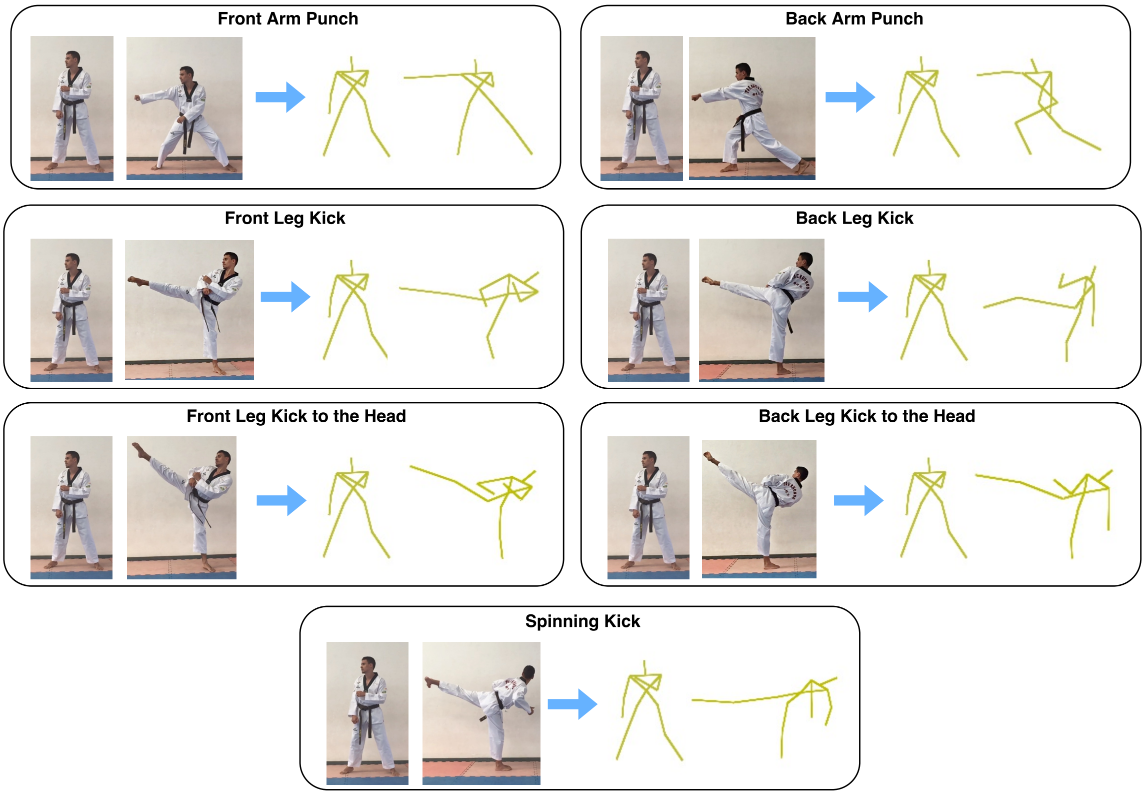


Figure 4.3: The seven moves and their corresponding extracted key poses.

provided by a decision tree but overcomes the overfitting drawback that complex trees have. In our experiment, in the training phase, we build a forest in which each root and node of a tree represents a key pose and the leaves are the moves, thus, a leaf-to-root path represents the possible sequence of key poses of that move[Faugeroux et al., 2014].

In the recognition process, when a new key pose is identified it is stored in a circular buffer and the tree rooted at that key pose is used to check, reversely iterating over the tree, if it completes a move. The construction of decision forest storing the gestures back-to-front simplifies the recognition work. Other benefit of this recognition process is that it works regardless the duration of the key pose performed by the athlete, the kind of input information we work in our experiments.

4.2 Dataset Assembly

As mentioned in the previous section, ten athletes where selected to perform the seven moves that are being studied in this research work, and in order to submit the moves execution information to the learning methods, the Taekwondo moves dataset was assembled.

The assembly process was made as follows: each athlete performed each move for 10 seconds, that is, the athlete 1 performed the front leg kick for 10 seconds, then performed the front leg kick to the head for 10 seconds, proceeding to the next moves until he completes the execution of all moves of the whole moves set, always taking a short pause between each move execution. Each move recorded is stored in a distinct file, until all athletes have performed all moves.

In the end of this process we have the total of 70 files of 10 seconds each, totalizing 11 minutes and 40 seconds of recording time. 1 minute and 40 seconds of recording time for each move, to be further submitted to validation, training and testing in both learning methods. Each move recording was stored in .CSV type files.

The information that is saved at every frame of the recording is composed of: 15 3D body joint coordinates, 9 zenith angles and 8 azimuth angles were recorded for each frame, forming a 62-element feature vector, as describe in Section 3.1. An example of a file of the dataset is shown in Apendix A.

Considering that we aimed in evaluating two kinds of features All-frames (unfiltered) and Key-poses, besides saving the information of all the frames of every move execution (unfiltered), our dataset also is composed of filtered files. Using the methodology described in the Section 3.2, for every move execution file that is crated we have a corresponding filtered file.

Summarizing, our dataset is composed of a total of 140 files, 70 of them being the unfiltered execution of each move by each athlete, and the remainder 70 are the corresponding Key-poses filtered files.

4.3 Moves Recognition

In this experiment, all the athletes performed each move separately for 10 seconds (i.e., the athlete 1 repeatedly performed the front leg kick for 10 seconds, took a pause, repeatedly performed the back leg kick for 10 seconds, took another pause and so on). Consequently, after the whole recording process, we captured 70 recordings of 10 seconds. In this case, we are evaluating the accuracy of our method, then we use these segmented sequences.

To evaluate the influence of key poses filtering step, we performed two different tests: i) All-frames: we disable the filtering step and execute the training/testing using All-frames; ii) Key-poses: tests with the key poses filtering on.

For each experiment (All-frames and Key-poses), we use 30% of each one for validating the LDCRF parameters (number of hidden states and regularization factor)

Experiment	Number of Hidden states	Regularization Factor
Key-poses	4	100.0
All-frames	5	0.0

Table 4.2: Best parameters found for our method using key pose filtering (Key-poses) and using the complete frame sequence (All-frames) in the validation phase. These values are used in all tests.

Methodology	Avg. Accuracy
Ours	74.72%
LDCRF + All-frames	58.29%
Decision Forest + Key-poses	51.02%
Decision Forest + All-frames	50.61%

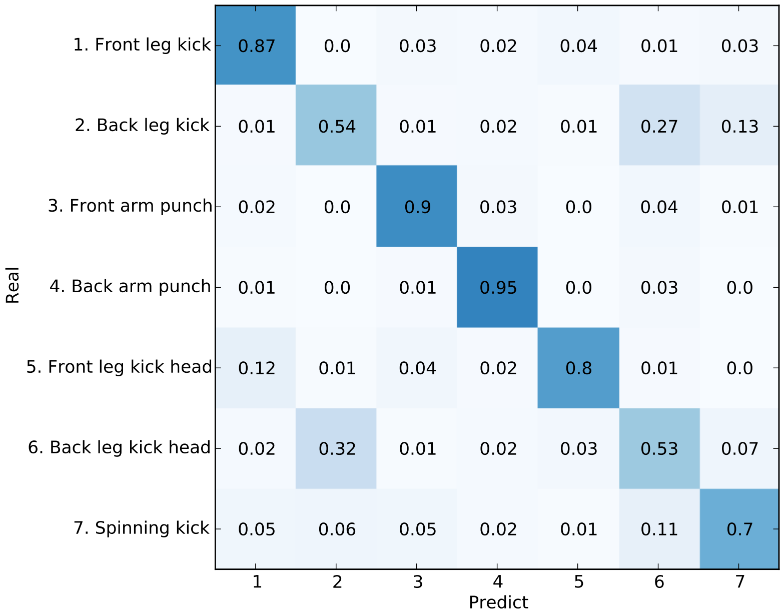
Table 4.3: Average accuracy of our methodology, LDCRF+All-frames and Decision Forest.

and the remaining 70% were used for training and tests. The best configuration for each experiment in the validation phase is summarized in Table 4.2: 5 hidden states and 0.0 regularization factor for All-frames; 4 hidden states and 100.0 regularization factor for Key-poses.

The number of hidden states parameter indicates the sum of hidden states that is associated to each class label of LDCRF chain. We varied or number of hidden states variable from 2 to 6 hidden states, according to the original variation that was tested in the LDCRF development. The regularization factor parameter determines the strength of the penalty on weight vectors whose norm is too large, in other words, this parameter tries to avoid the overfitting of the model generated, that is the situation where the model generated is too specific for the trained parameters. Ideally overfitting is not desirable in learning methods since it makes the model work well on the trained parameters and not so well on new sample submitted to it.

Using the remaining 70% of the whole dataset and the best parameters values discovered in the validation phase, we trained and tested our model using a K-fold cross validation method. Our dataset was divided into 5 subsets, and the training and testing method was repeated 5 times. Each time, one of the 5 subsets is used as the test set and the other 4 subsets are put together to form a training set. Then the average recognition rate across all 5 trials is computed.

As the results show in Table 4.3, our filtering methodology using Key-poses yield



	1	2	3	4	5	6	7
1. Front leg kick	0.87	0.0	0.03	0.02	0.04	0.01	0.03
2. Back leg kick	0.01	0.54	0.01	0.02	0.01	0.27	0.13
3. Front arm punch	0.02	0.0	0.9	0.03	0.0	0.04	0.01
4. Back arm punch	0.01	0.0	0.01	0.95	0.0	0.03	0.0
5. Front leg kick head	0.12	0.01	0.04	0.02	0.8	0.01	0.0
6. Back leg kick head	0.02	0.32	0.01	0.02	0.03	0.53	0.07
7. Spinning kick	0.05	0.06	0.05	0.02	0.01	0.11	0.7

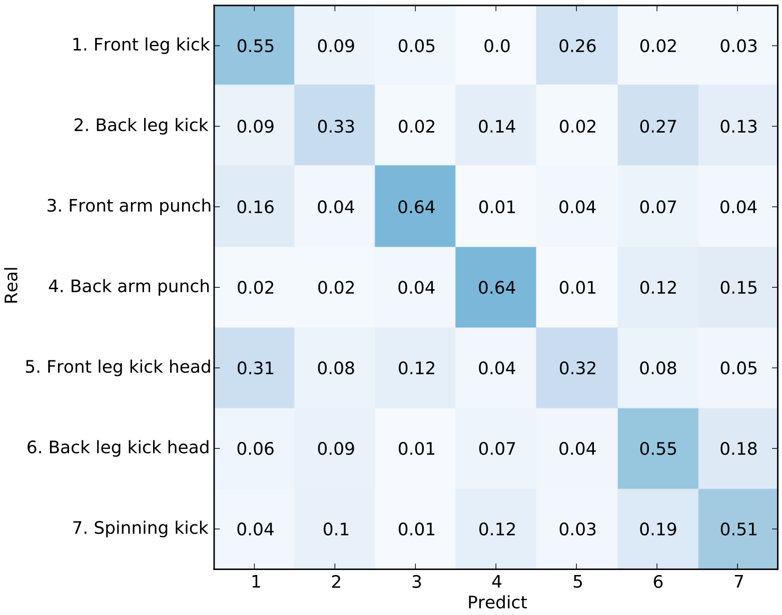
Figure 4.4: Confusion matrix of our methodology.

the best results when compared with the use of All-frames. This is the result of feeding the recognition process with several contiguous frames that are very similar among them. Our methodology works using the LDCRF learning method, and as it works mapping the dynamics between frames, confirming our initial hypotheses that these redundant frames harm the modeling process.

We compare the performance of our methodology against the Decision Forest adopted in Faugeroux et al. [2014], which also applies a filtering step to select the more discriminative frames. The average recognition rates obtained are all summarized in Table 4.3. One can clearly see that our recognition methodology outperforms the Decision Forest approach in both scenarios, presenting a considerably higher accuracy.

Another way to analyse results of machine learning techniques is to create a confusion matrix. Each column of the matrix represents the instances of the predicted class while each row represents the instances that where actually recognized by the method. The advantage of generating a confusion matrix is that it provides a clear visualization of the strong and weak point of the model generated. The perfect confusion matrix is the one that shows a diagonal matrix with all values 1.0. This situation is the ideal one in computer vision systems, since it indicates that the model generated is the one that was able to map the particularities of each class correctly.

The confusion matrix of our approach is shown in Figure 4.4. It shows very good results in front leg kick, front leg kick to the head, front arm punch, back arm punch



Real	1	2	3	4	5	6	7
1. Front leg kick	0.55	0.09	0.05	0.0	0.26	0.02	0.03
2. Back leg kick	0.09	0.33	0.02	0.14	0.02	0.27	0.13
3. Front arm punch	0.16	0.04	0.64	0.01	0.04	0.07	0.04
4. Back arm punch	0.02	0.02	0.04	0.64	0.01	0.12	0.15
5. Front leg kick head	0.31	0.08	0.12	0.04	0.32	0.08	0.05
6. Back leg kick head	0.06	0.09	0.01	0.07	0.04	0.55	0.18
7. Spinning kick	0.04	0.1	0.01	0.12	0.03	0.19	0.51

Figure 4.5: Confusion matrix of the Decision Forest method.

and spinning kick. The spinning kick is a one of a kind move, since it differs a lot from all other moves, being easier for the methodology to learn its unique characteristics. The kicks performed with the front leg (at chest height and at the head height) have a smaller body rotation with consequent less joint shifting and better recognition results. The punches are the simplest moves of our moves set, small body rotation and the predominance of one arm dislocation, where we expected the good results shown in the confusion matrix.

The performance degradation seen in the two back leg kicks happens due to the complexity and similarity of them. They are as similar between them as the front leg kicks are between them, but the back leg kicks' bigger body rotation, with the consequent bigger joint shifting, makes them more complex and more difficult for the learning method.

The problem of moves with big joint shifting concerns the Kinect video capture and body estimations technology limitations. Kinect works estimating each body joint of a person in the 3D space coordinate. In complex moves, high speed and big joint shifting, the imprecision of the Kinect increases and it is propagated to the data that is submitted to the recognition method, harming the feature vector that is stored at each frame of the video recording, making it hard for the model to map the transitions between one frame and its neighbors.

Figure 4.5 shows the confusion matrix of the Decision Forest approach. When

Dataset	Running time (s)
Key-poses	10.5
All-frames	78.5

Table 4.4: Running time comparison in All-frames and Key-poses experiments.

comparing the confusion matrix of our approach and the confusion matrix of the Decision Forest, we can clearly see that our methodology outperforms the Decision Forest recognition accuracy in all kinds of moves individually, except by the back leg kick to the head, where both methods achieved similar results.

The confusion matrix of the Decision Forest approach also shows that this methodology is not able to generate a model that can precisely differentiate both front leg kick moves and back leg kick moves. This matrix shows that a high percentage of back leg kicks are recognized as back leg kicks to the head, and a high percentage of front leg kicks to the head are recognized as front leg kicks, that is, a huge recognition impact on moves with high similarity between them (back leg kick/back leg kick to the head and front leg kick/front leg kick to the head). In addition we can see a huge recognition impact on the moves with high body rotation (back leg kicks and spinning kick).

Both confusion matrices show a better accuracy in punch moves. Like we said before, these moves are the ones that have smaller body rotation, which makes of them less complex moves when comparing to the kick moves, allowing the learning method to generate better models for these moves.

We also evaluate the running time of our methodology. As expected, Table 4.4 shows that our approach performs a routine recognition faster when using Key-poses. The running times exhibited in the Table 4.4 correspond to the processing and recognition of short training session: 280 moves, resulting in the total of 371 frames for the Key-poses dataset and 2940 frames for the All-frames dataset.

In long training sessions executed by the athletes, the seven times faster execution time of our methodology will show to be even more relevant. A faster running time also allows the system to perform an on-line recognition and segmentation, when needed.

4.4 Temporal segmentation

The existence of a system that performs automatic recognition and temporal segmentation of an athlete training routine is unknown for the regular market, according to the high experienced Taekwondo coach that conducted our experiments in the CTE-

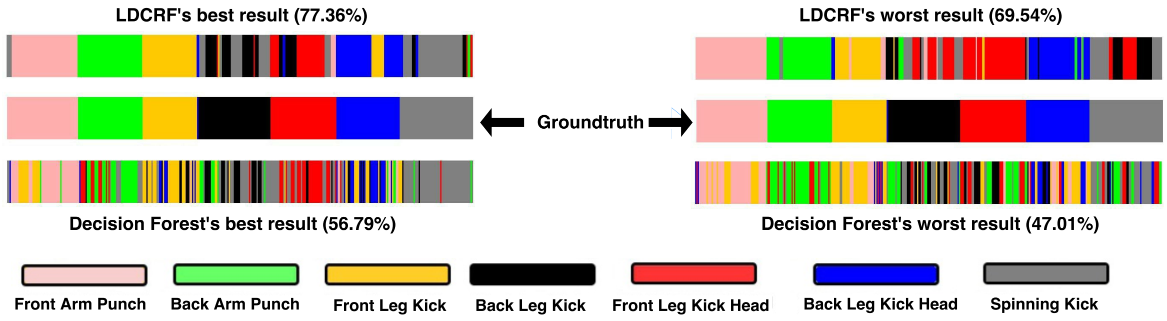


Figure 4.6: Two sequence segmentation charts for our methodology and Decision Forest.

UFMG. This kind of system is usually tailored designed and restricted to few training centers.

Temporal segmentation means partitioning a recorded video according to the action that is performed in a certain slot of time. It involves two important quality matters: i) the correct location of the point in time where the actions change; ii) the correct duration of a specific movement execution.

In the segmentation experiments we performed, we used the methods combined with Key-poses filtering, since it was the one that produced the best result in both recognition techniques. Our goal in this experiment is not only to evaluate the accuracy of the recognition, but also compare the quality of the segmentation of our technique with the Decision Forest approach.

To evaluate the temporal segmentation of both methodologies, we submitted the learning methods to 10 new training routine sequence samples. The total time of each sequence is of 1 minute and 10 seconds and we used the key poses as input for the learning methods.

Figure 4.6 presents the best and the worst segmentation provided by both methodologies as well as the segmentation groundtruth for the sequence. In the perfect scenario, groundtruth, the quality of the temporal segmentation achieves 100% in both duration of the move and time where moves change. In our methodology, that uses a discriminative model, we achieve 77.36% and 69.54% as our best and worst accuracies for segmentation, while Decision Forest obtained are 56.79% and 47.01%, respectively.

The charts show that, as a result of a higher recognition rate, our methodology produces a more smooth and accurate segmentation chart than Decision Forest. We can also notice that our method exhibit a bad segmentation result for the back leg kick, that is coherent to the fact it is one of the most difficult moves to be recognized, due to a huge body spin combined with the high elevation of the leg. In less complex moves like the punches, our methodology shows high accuracy temporal segmentation.

The Decision Forest charts show high imprecision in temporal segmentation blocks even in its best result. In great part of both Decision Forest charts it is not possible to identify precisely where the transition between moves occur. This fact prevents this method from being used for temporal segmentation. On the contrary, in our methodology we can clearly see the moves transitions, although with some natural imprecision due to the errors in the recognition method. The results presented by our methodology are very good results, considering the high complexity of the moves being performed and the lack of a similar system.

Chapter 5

Conclusion

In this dissertation we presented a methodology to recognize moves and segment sequences in high performance sports moves based on key poses. We evaluated our methodology on top of Taekwondo, an Olympic sport where athletes launch strokes as fast as possible.

Our approach, that combines the LDCRF learning method and Key-poses filtering, proved to be not only more efficient, but it also achieved both a higher recognition rate and a more accurate segmentation compared to a state-of-the-art technique. In particular, it achieved a 74.72% recognition accuracy using only 13% of the frames from a 30 FPS video training routine, compared to the 58.29% of accuracy of using all frames of the video training routine. The advantages of using the key poses method are not only restricted to recognition rate and segmentation, i.e., the amount of data produced and time required are also reduced. Our results also show that our methodology yields better recognition rate when comparing to the Decision Forest, in both Key-poses and All-frames experiments.

The LDCRF was initially designed to receive all sequential data of a event to create its recognition model, however, in our challenging scenario it showed that it can be used efficiently with only some key part of a sequential event. In fact, the redundant information provided by a high frame rate did harm the model created by the LDCRF, confirming our initial hypotheses.

Finally, it is worth to note that our methodology uses inexpensive equipment and does not required a strictly controlled environment to give good recognition results. In addition, it is generic enough to be applied not only to Taekwondo, but also to a large range of other high performance sports, such as Boxing, Fencing, Tennis, Golf and Archery.

The sports previously mentioned are some examples of sports in which the athlete can do his training individually, since our methodology was initially designed to work with one person. However, with upgrades and new algorithms, it can be used with more than one person in the scene and not only to recognize the movements but also to assess the quality of the movement performed when compared to the execution of an expert.

Chapter 6

Future Works

In this dissertation we worked not only on recognizing and segmenting Taekwondo moves but also on assessing the Kinect quality as a reliable move time measurement device. The time measurement experiments were made independently of the recognition and segmentation, and these time experiments were used as preliminary tests for moves recognition. One natural evolution of this work is to combine both modules: the moves recognition and the kick time measurement, therefore, at the same time we would identify the kick moves and measure their execution time.

Concerning the recognition method comparison, we can compare our methodology (LDCRF + Key+poses) with other recognition methods that also map the temporal relation among samples. One possible candidate is the Hidden Markov Model (HMM) method. Like the LDCRF the HMM also works creating the unobserved variable (hidden variables) that map the transitions among samples.

Another clear upgrade to this research work would be to add not only another person in the scene but also a greater variety of moves. Adding another person is important because in part of his training session, an athlete trains with another opponent to develop his moves, furthermore one coach usually helps the athlete with a racket or other Taekwondo apparel. The movement of the coach's body does not have to be recognized by the system, but its presence in the scene interferes in the recognition of the athletes moves. The addition of a greater variety of moves would make possible for an athlete to use all his skills in our methodology.

In addition, we can also evaluate the recognition performance of our methodology using the frontal perspective. Although the frontal perspective would not be adequate to a scenario where two athletes are fighting each other, or the scenario where the athlete is training with a coach it would be useful in the scenario where an athlete is training alone in order to develop his technique. The problem of this perspective, when we have two or more persons in the scene, occurs due to occlusion, where the joints of the person are not visible by the sensors harming the pose estimation and the recognition process.

The occlusion problem can be overcome with the use of more than one Kinect sensor, which has to be done carefully. In preliminary tests done by us, we noticed that using more than one Kinect can lead to sensor interference, resulting in poor skeleton mapping in one or all sensors in the scene. We can use more than one sensor to track an athlete's move, but we have to discard poor quality joint mappings, which can be done by using the joint confidence parameter provided by the Kinect.

An evaluation that is also necessary in the filtering step of our methodology is concerning what frame, of the contiguous frames that have the same key pose label associated to them in the filtering process, shall we choose. In our experiments we chose the first one, but further evaluations are necessary in order to find what is the best sample among all the contiguous frames with the same label. Choosing the middle frame is one example of the tests that shall be done.

Finally, we could apply our methodology in other high performance Olympic sports like Boxe and Fancing.

Bibliography

Bialkowski, A., Lucey, P., Carr, P., Denman, S., Matthews, I., and Sridharan, S. (2013). Recognising team activities from noisy data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 984–990.

Bianco, S. and Tisato, F. (2013). Karate moves recognition from skeletal motion. In *IS&T/SPIE International Symposium on Electronic Imaging*, pages 86500K--86500K. International Society for Optics and Photonics.

Borras, J. and Asfour, T. (2015). A whole-body pose taxonomy for loco-manipulation tasks. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 1578–1585.

Carfagno, D. G. and Hendrix, J. C. (2014). Overtraining syndrome in the athlete: Current clinical practice. *Current Sports Medicine Reports*, 13(1):45--51.

Catuhe, D. (2012). *Programming with the Kinect for Windows Software Development Kit (Developer Reference)*.

Choppin, S. and Wheat, J. (2013). The potential of the microsoft kinect in sports analysis and biomechanics. *Sports Technology*, 6(2):78–85.

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., and Del Bimbo, A. (2013). Space-time pose representation for 3d human action recognition. In *New Trends in Image Analysis and Processing ICIAP 2013*, Lecture Notes in Computer Science, pages 456--464.

El-Sallam, A., Bennamoun, M., Sohel, F., Alderson, J., Lyttle, A., and Rossi, M. (2013). A low cost 3D markerless system for the reconstruction of athletic techniques. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 222–229. ISSN 1550-5790.

Elgendi, M., Picon, F., and Magenant-Thalmann, N. (2012). Real-time speed detection of hand gesture using kinect. In *Workshop on Autonomous Social Robots and Virtual Humans - 25th Annual Conference on Computer Animation and Social Agents (CASA)*, pages 09--11.

Faugeroux, R., Vieira, T., Martinez, D., and Lewiner, T. (2014). Simplified training for gesture recognition. In *27th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 133–140.

Gade, R. and Moeslund, T. (2013). Sports type classification using signature heatmaps. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 999–1004.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73:82--98.

Ge, S. and Fan, G. (2015). Non-rigid articulated point set registration for human pose estimation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 94–101.

Gong, D., Medioni, G., and Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 36, pages 1414–1427. ISSN 0162-8828.

Gong, D., Medioni, G., Zhu, S., and Zhao, X. (2012). Kernelized temporal cut for online temporal segmentation and recognition. In *12th European Conference on Computer Vision (ECCV)*, pages 229--243. Springer-Verlag.

Gray, A. D., Marks, J. M., Stone, E. E., Butler, M. C., Skubic, M., and Sherman, S. L. (2014). Validation of the microsoft kinect as a portable and inexpensive screening tool for identifying acl injury risk. *Orthopaedic Journal of Sports Medicine*.

Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. *Sports Medicine*, 4(2):139--147.

Hamid, R., Kumar, R., Hodgins, J., and Essa, I. (2014). A visualization framework for team sports captured using multiple static cameras. *Computer Vision and Image Understanding*, 118:171--183.

Hu, R. M., He, Z. D., and Bai, F. (2014). The research of 3D human motion simulation and video analysis system implemented in sports training. *Advanced Materials Research*, pages 2743--2746.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using hausdorff distance. *IEEE Transaction on Pattern Analysis and Machine Intelligence*.

Jarvie, G. (2012). *Sport, Culture and Society: An Introduction*. Routledge, second edition.

k. Ho, T. (1995). Random decision forests. *Third International Conference on Document Analysis and Recognition*, pages 278--282.

Keegan, R. J., Harwood, C. G., Spray, C. M., and Lavallee, D. (2014). A qualitative investigation of the motivational climate in elite sport. *Psychology of Sport and Exercise*, 15:97--107.

Kirk, A., O'Brien, J., and Forsyth, D. (2005). Skeletal parameter estimation from optical motion capture data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 782–788 vol. 2. ISSN 1063-6919.

LaViola, J. (2013). 3D Gestural Interaction: The State of the Field. *ISRN Artificial Intelligence*, page 514641.

Lea, C., Hager, G., and Vidal, R. (2015). An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1123–1129.

Mailapalli, D. R., Benton, J., and Woodward, T. W. (2015). Biomechanics of the taekwondo axe kick: a review. *Journal of Human Sport & Exercise*, 10:141--149.

Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A. W., and Campos, M. F. M. (2012). Real-time gesture recognition from depth data through key poses learning and decision forests. In *25th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 268--275.

Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A. W., and Campos, M. F. M. (2014). Online gesture recognition from pose kernel learning and decision forests. *Pattern Recognition Letters*, 39:65--73.

Morency, L., Quattoni, A., and Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. ISSN 1063-6919.

Pers, J. and Kovacic, S. (2000). Computer vision system for tracking players in sports games. In *1st International Workshop on Image and Signal Processing and Analysis (IWISPA)*, pages 177–182.

Shi, Q., Wang, L., Cheng, L., and Smola, A. (2008). Discriminative human action segmentation and recognition using semi-markov model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. ISSN 1063-6919.

Spriggs, E., De la Torre, F., and Hebert, M. (2009). Temporal segmentation and activity classification from first-person sensing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 17–24. ISSN 2160-7508.

Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., and Campos, M. F. M. (2013). On the improvement of human action recognition from depth map sequences using spacetime occupancy patterns. *Pattern Recognition Letters*, 36:221--227.

Wada, S., Fukase, M., Nakanishi, Y., and Tatsuta, L. (2013). In search of a usability of kinect in the training of traditional japanese “KATA” stylized gestures and movements. In *2nd International Conference on e-Learning and e-Technologies in Education (ICEEE)*, pages 176–179.

Wang, Y. and Mori, G. (2009). Max-margin hidden conditional random fields for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 872–879. ISSN 1063-6919.

Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., and Rehg, J. M. (2007). A scalable approach to activity recognition based on object use. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8. ISSN 1550-5499.

Xia, L., Chen, C.-C., and Aggarwal, J. (2012). View invariant human action recognition using histograms of 3D joints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20--27.

Yu, G., Liu, Z., and Yuan, J. (2014). Discriminative orderlet mining for real-time recognition of human-object interaction. In *12th Asian Conference on Computer Vision (ACCV)*, Lecture Notes in Computer Science, pages 50--65.

Zecha, D. and Lienhart, R. (2015). Key-pose prediction in cyclic human motion. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 86–93.

Appendix A

Dataset Sample File

In this Apendix we show a sample file of the dataset. Each file is a .csv file which contains the execution data of a move performed by one athlete. The first column of the file shows the type of information of each line of the file. The first 17 lines of each file contain the Joint-angle information of the athlete body and the 45 remainder lines contain the 3D coordinate of each Joint, as describe in Section 3.1.

Each column of the file contain the information (Joint-angles and Joints) of the athlete's body at each frame. If we are performing an All-frames experiment, the number of columns recorded in the file will correspond to all frames of the recording. In the case of a Key-poses experiment the number of columns recorded will correspond to the filtered key-poses frames.

The table below shows an excerpt of a front leg kick move.

TYPE OF INFORMATION	FRAME DATA			
	1	2	...	n
Raised Right Arm Zenith Angle	18.044744	23.404295	...	41.087326
Raised Left Arm Zenith Angle	23.656776	26.037018	...	25.478489
Opened Right Arm Azimuth Angle	80.988022	86.796616	...	-122.556488
Opened Left Arm Azimuth Angle	74.615639	75.155731	...	69.379494
Raised Right Leg Zenith Angle	34.677326	40.462528	...	44.055847
Raised Left Leg Zenith Angle	38.130394	45.079510	...	48.911682
Opened Right Leg Azimuth Angle	-79.760857	-92.459747	...	-46.710373
Opened Left Leg Azimuth Angle	115.890961	116.967018	...	118.695808
Right Elbow Zenith Angle	19.079357	26.712439	...	16.344690
Right Elbow Azimuth Angle	-71.048759	-88.154121	...	-74.982933
Left Elbow Zenith Angle	127.038757	125.119110	...	133.749695

Left Elbow Azimuth Angle	53.861179	50.085880	...	54.745392
Right Knee Zenith Angle	45.516853	32.128170	...	71.018501
Right Knee Azimuth Angle	44.526173	-57.936901	...	-27.354166
Left Knee Zenith Angle	-30.833824	27.992460	...	18.889299
Left Knee Azimuth Angle	-92.611557	-87.250565	...	-84.815453
Japanese Bow Zenith Angle	25.890375	26.931307	...	30.478579
Left Shoulder X Coordinate	-799.444092	-811.584534	...	-813.262024
Left Shoulder Y Coordinate	70.626663	40.415947	...	67.647972
Left Shoulder Z Coordinate	2359.362793	2361.351318	...	2349.046875
Right Shoulder X Coordinate	-614.434448	-644.377014	...	-652.025024
Right Shoulder Y Coordinate	223.162811	209.411835	...	231.381287
Right Shoulder Z Coordinate	2357.675781	2347.212158	...	2329.121826
Head X Coordinate	-835.837646	-872.172974	...	-879.683350
Head Y Coordinate	303.974762	268.179657	...	290.152893
Head Z Coordinate	2303.776123	2292.661133	...	2279.845215
Torso X Coordinate	-575.202026	-581.051758	...	-582.555298
Torso Y Coordinate	-7.347495	-15.020628	...	13.344205
Torso Z Coordinate	2365.005127	2366.620605	...	2351.27124
Left Hip X Coordinate	-375.174561	-373.816040	...	-375.654907
Left Hip Y Coordinate	-105.286041	-94.003448	...	-65.134491
Left Hip Z Coordinate	2370.986084	2374.630615	...	2358.296387
Right Hip X Coordinate	-511.755005	-494.429443	...	-489.279175
Right Hip Y Coordinate	-217.893417	-215.906860	...	-180.517944
Right Hip Z Coordinate	2371.99585	2383.28833	...	2370.506348
Left Elbow X Coordinate	-701.584167	-702.652832	...	-713.580933
Left Elbow Y Coordinate	-76.187454	-80.509720	...	-68.835869
Left Elbow Z Coordinate	2367.018066	2366.695068	...	2370.572021
Right Elbow X Coordinate	-393.333313	-391.546997	...	-620.175171
Right Elbow Y Coordinate	72.626587	66.421570	...	-32.938599
Right Elbow Z Coordinate	2316.508545	2321.187256	...	2314.43457
Left Hand X Coordinate	-607.800293	-637.140564	...	-667.829712
Left Hand Y Coordinate	188.480453	191.397812	...	211.881622
Left Hand Z Coordinate	2343.697266	2334.627686	...	2325.235596
Right Hand X Coordinate	-187.091156	-144.714722	...	-515.388672
Right Hand Y Coordinate	-32.183647	60.243599	...	-277.641968

Right Hand Z Coordinate	2156.663818	2153.818115	...	2148.324219
Left Knee X Coordinate	-482.521729	-478.389679	...	-492.461670
Left Knee Y Coordinate	-617.402954	-618.879639	...	-586.185059
Left Knee Z Coordinate	2222.288574	2212.428711	...	2203.641113
Right Knee X Coordinate	-288.277527	-320.139679	...	-126.514374
Right Knee Y Coordinate	-463.538300	-480.872986	...	-166.454742
Right Knee Z Coordinate	2202.699463	2201.449707	...	2204.034912
Left Foot X Coordinate	-617.141968	-631.159790	...	-604.051758
Left Foot Y Coordinate	-962.132690	-972.953979	...	-915.041321
Left Foot Z Coordinate	2206.77832	2197.165039	...	2188.632324
Right Foot X Coordinate	-303.005981	-328.656036	...	130.782074
Right Foot Y Coordinate	-836.285278	-859.991455	...	-397.181824
Right Foot Z Coordinate	2392.678711	2379.994141	...	2380.120605
Neck X Coordinate	-706.939270	-727.980774	...	-732.643555
Neck Y Coordinate	146.894745	124.913895	...	149.514633
Neck Z Coordinate	2357.974121	2388.546631	...	2398.350586

Table A.1: Sample of a dataset move file.