sensors

MDPI

Article

Temporal Pattern Attention for Multivariate Time Series of Tennis Strokes Classification

Maria Skublewska-Paszkowska *,† **and Pawel Powroznik** †

Department of Computer Science, Lublin University of Technology, 20-618 Lublin, Poland
* Correspondence: maria.paszkowska@pollub.pl
† These authors contributed equally to this work.

Abstract: Human Action Recognition is a challenging task used in many applications. It interacts with many aspects of Computer Vision, Machine Learning, Deep Learning and Image Processing in order to understand human behaviours as well as identify them. It makes a significant contribution to sport analysis, by indicating players' performance level and training evaluation. The main purpose of this study is to investigate how the content of three-dimensional data influences on classification accuracy of four basic tennis strokes: forehand, backhand, volley forehand, and volley backhand. An entire player's silhouette and its combination with a tennis racket were taken into consideration as input to the classifier. Three-dimensional data were recorded using the motion capture system (Vicon Oxford, UK). The Plug-in Gait model consisting of 39 retro-reflective markers was used for the player's body acquisition. A seven-marker model was created for tennis racket capturing. The racket is represented in the form of a rigid body; therefore, all points associated with it changed their coordinates simultaneously. The Attention Temporal Graph Convolutional Network was applied for these sophisticated data. The highest accuracy, up to 93%, was achieved for the data of the whole player's silhouette together with a tennis racket. The obtained results indicated that for dynamic movements, such as tennis strokes, it is necessary to analyze the position of the whole body of the player as well as the racket position.

Keywords: sport; tennis strokes; human action recognition; A3T-GCN; motion capture

Citation: Skublewska-Paszkowska, M.; Powroznik, P. Temporal Pattern Attention for Multivariate Time Series of Tennis Strokes Classification. *Sensors* **2023**, *23*, 2422. https://doi.org/10.3390/s23052422

Academic Editors: Miguel Correia and Leandro José Rodrigues Machado

Received: 22 December 2022
Revised: 7 February 2023
Accepted: 20 February 2023
Published: 22 February 2023

1. Introduction

Computer Vision is an interdisciplinary field of study that aims to derive meaningful information from various types of data. Applying artificial intelligence for digital images, skeleton, depth, videos, point cloud, audio, acceleration, signals or motion capture data allows one to perform actions or make decisions as well as further recommendations. The purpose of Human Action Recognition (HAR) is to understand human behaviours and identify them [1,2]. It specifies a set of person's moves performed in time in order to complete a task. Occasionally, additional objects, such as a tennis racket or a golf club, are involved to do the actions. Depending on the complexity of the movements and their duration, different length sequences are taken into consideration, from a single frame to a whole video streaming. HAR is a challenging task used in numerous applications. It interacts with many aspects of Computer Vision, Machine Learning, Deep Learning and Image Processing [3]. It utilizes detection of a person or objects in the image, video as well as sensor data, the location of the action in time and space, and the recognition of the action. This attitude usually involves feature detection, such as extracted from 3D silhouettes, skeletal joint and body part location, local spatio-temporal, local occupancy patterns and finally 3D scene flow [2]. That is why it makes a significant contribution to sport analysis. Detection of athletes and recognition of their actions or teams' activities plays a pivotal role in indicating the players' performance level and training evaluation or analyzing sport statistics [3].

Classification is a challenging task; however, its use can be found in many studies considering various sport disciplines, for both individual and team ones. Twelve basketball activities, corresponding to the most of the complex actions, were recognised in [4]. For this purpose the LSTM-DGCN method was proposed. It consisted of two parts: Deep Graph Convolutional Network (DGCN) and Long Short-Term Memory (LSTM). Basketball players measures such as distances between joints and angles were input parameters. Moreover, selection coordinates and depth maps together with RGB frame sequences were used for this purpose as well.

Both action recognition and analysis of the karate athletes were presented in [5]. The Attention-enhanced Graph Convolutional LSTM Networks (AGC-LSTM) was applied for recognition of the actions gathered from five athletes. The analyzed movements such as a punch (middle, upper, forward upper and back upper), back pounce, kick (side rising, back, inside crescent, side) as well as several technical moves were taken into consideration. A framework dedicated to sports video recording considering attentive movement characterization was presented in [6]. It involved hierarchical recurrent neural networks. The extraction of human pose, use of trajectory clustering made it possible to describe a dynamic movement of players or the whole teams as well as various interactions in sport games, such as volleyball. The classification of water sports using the Convolutional Neural Network (CNN) with discriminative filter banks was presented in [7]. Water skiing and surfing were taken into consideration. Both 2D and 3D data were analysed. Three branches were applied for the classifier consisting of: the average pooling classification head, a set of convolutions, spatial upsampling and max-pooling layers.

Recently, a novel approach to analyzing the movements of athletes has been proposed. It took into consideration the natural connections between the joints of the human silhouette in order to apply the graphs. The use of Graph Convolutional Neural Networks (GNN) made it possible to capture existing patterns and dependencies embedded in a spatial configuration of joints, as well as their temporal dynamics and thus it has become a very popular method in the field of HAR [8]. In that study, the authors introduced the Spatial Temporal Graph Convolutional Networks (ST-GCN) for everyday activities, sports areas and actions recognition. The same network was applied for the recognition of the selected movements in competitive sports [9]. Activities such as balance beam, diving, athletics, boxing, keeping fit, and badminton were taken into consideration. The data were recorded with the use of the markerless motion capture system. In [10], it was stated that a model based on temporal convolutional networks was more appropriate for HAR than a model considering the recurrent neural ones. These two approaches were compared using the NTU RGB+D dataset containing various types of actions. In [11], a new Two-stream Adaptive Graph Convolutional Network (2s-AGCN) for skeleton-based action recognition was proposed, which was verified by two datasets NTU-RGBD and KineticsSkeleton. This approach included additional information of skeleton data, such as the bone location. This attitude enhanced the performance of the classifier. The ST-GCN for skeleton action recognition was also applied for a similar approach in [12]. The Actional-Structural Graph Convolution Network (AS-GCN) was proposed. Its structure was characterised by basic building blocks for indicating spatial and temporal features. This new classifier was verified with NTU-RGBD+D and Kinetics datasets. Various types of sports, such as: golf, kicking, lifting, diving, running, horse riding, skateboarding, swing-bench, and walking, were recognized using the Part-Attention Spatio-temporal Graph Convolutional Network (PSGCN) [13]. It exploited the dynamic information from a sports video.

Classification of the tennis movement may be found in many scientific papers. The studies involved only skeleton-based action recognition, only tennis racket position as well as the whole player silhouette together with a racket. The research concerning SensorTile attached to the tennis racket was presented in [14,15]. For the purpose of swing classification Deep Neural Network [14] was applied, while for topspins (forehand and backhand), subpar forehand, subpar backhand, and slices (forehand and backhand) recognition the following methods were used: Support Vector Machine (SVM), Neural Networks (NN),

Decision Tree (DT), Random Forest (RF) and k-Nearest Neighbor (kNN) [15]. The Pan Tompkins algorithm for the classification of shots using time warping was presented in [16]. Many studies focused on tennis stroke recognition based on video data. This attitude involved extracting features from videos and applying a classifier to the whole set [17]. The THETIS is a very well-known dataset consisting of twelve tennis moves captured by Microsoft Kinect in a form of video and ONI files [18]. The video-based action recognition of backhand (two-handed, one-handed, slice, and volley), forehand (flat, open stance, slice, and volley), serve (flat, kick, and slice) as well as smash was performed using the 3-layered LSTM network in [19,20]. In [17], these twelve moves were classified using SVM and linear-chain Conditional Random Fields (CRF). The five-layer deep historical LSTM network described in [21] was applied for similar moves using the following datasets: THETIS and HMDB51. Six tennis strokes from the THETIS datasets were recognised by the LSTM network in [22]. Serve, hit as well as non-hit were recognized by the Kernelised Linear Discriminant Analysis (KLDA) in [23]. Transductive transfer learning for an annotation of video sequences was applied. The changes in the tennis ball were also taken into consideration. The basic tennis strokes, forehand and backhand, from a video were analyzed in [24–26] using the SVM classifier. In [27], tennis serves, forehand and backhand were recognised using two classifiers: SVM with the radial basis function kernel and K-Nearest Neighbour classifiers (KNNs). A wireless inertial measurement unit sensor together with a system consisting of eight video cameras was used for capturing the data.

Studies concerning HAR were also performed using motion capture data recorded via optical systems. Forehand and backhand strokes with and without ball contact as well as no-shots were recognized by ST-GCN based on images generated from three-dimensional motion data in [28]. Graph Convolutional Networks (GCNs) were an obvious choice due to the fact that the parts of the image correlated with the human topology. In this study, the influence of input fuzzification on the obtained accuracy was examined. The results showed that this approach increased recognition ability. An extension of the above research was the recognition of individual tennis stroke phases, i.e., forehand preparation, forehand shot with racket swinging, backhand preparation, backhand shot with racket swinging and no-shots which were presented in [29]. Three classifiers with and without fuzzification were taken into account: SVM, MLP, and ST-GCN. In addition, the influence of the extensions and generalizations of the Choquet integral on the aggregation of results obtained by individual classifiers was verified. The results indicated that this method increased the efficiency of recognizing tennis moves. Another approach to tennis movements recognition including its phases was presented in [30]. For the purpose of the classification, the Attention Temporal Graph Convolutional Network (A3T-GCN) was applied both with and without input fuzzification. The conducted results showed that this classifier might be considered as one of the most appropriate methods for tennis classification.

The state-of-the-art study presented in this paper is to apply the A3T-GCN classifier for tennis stroke recognition based on three-dimensional coordinates data obtained from the optical motion capture system. Forehand, backhand and volley strokes were taken into consideration. The main purpose of this study is to look into how the content of three-dimensional data influences classification accuracy, precision, recall, and F1 score. Both the coordinates associated with the player's silhouette and the position of the racket were analyzed, which to the authors' knowledge is the novelty approach. The A3T-GCN was chosen due to the attention model, which both stores information about the player's model, but also determines the predicted player position.

The rest of this paper is organized as follows. Section 2 explores the material and methods as well as introduces the Attention Temporal Graph Convolutional Network. Section 3 presents results of the state-of-the-art action recognition methods with the proposed classifier and 3D motion capture data. Section 4 discusses the proposed method, and finally Section 5 concludes the study.

2. Materials and Methods

2.1. Participants

In this study, seven male and three female tennis players took part (age 23.7 ± 4.58, height 1.77 ± 0.13 m, weight 71.65 ± 10.68 kg). Only one of them was left-handed, while the others were right-handed. They all signed the consent for the study.

2.2. Data Acquisition

Each participant was prepared for the experiment. First, they have a 15-min warm-up. Second, thirty-nine retroreflective markers, specified in the Plug-in Gait model, were attached to their body. Finally, all the required measurements were gathered for the purpose of creating a new model as well as preparing its calibration in the motion capture system. Furthermore, seven markers were also attached to the tennis racket, according to the following scheme: one to the top of the racket head, two on both sides of the racket, one to the bottom of the racket head and one to the bottom of the racket handle. Such an arrangement reflects the racket shape and capture its movements.

For the purpose of acquisition, eight-camera optical Vicon motion capture system, installed in the indoor room, was used with the Nexus software. The cameras are mounted two on each wall on the same level. The whole schema of the cameras arrangement is presented in Figure 1. Before movement acquisition the calibration of the system was performed. The maximal calibration error did not exceed 0.045 pixels. The frequency of capturing was set to 100 Hz.

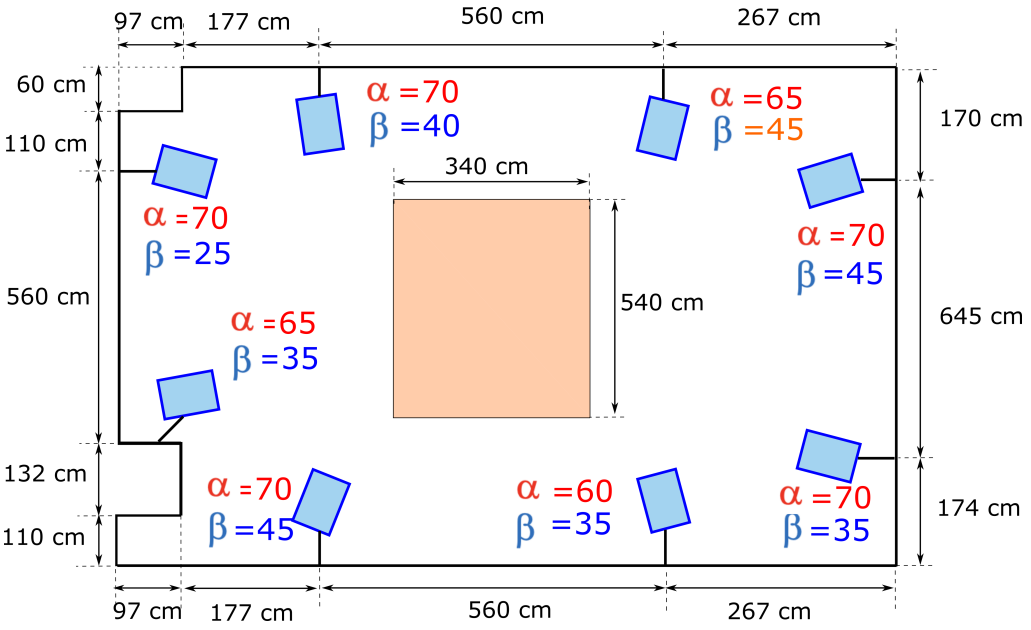


Figure 1. Motion capture cameras arrangement, where α is the angle in the OX plane between the floor and the camera axis, β is the angle in the OY plane between the camera axis perpendicular to the floor and the camera.

Each participant performed forehand, two-handed backhand and volley strokes. Forehand and backhand ones were performed while running and avoiding a bollard placed on the floor. Due to this, the strokes were more natural than hitting the ball from a standing position. At first, ten forehand strokes without a ball were performed, followed by ten backhand strokes without a ball. Next, these exercises were repeated with a ball. Finally, the participant performed ten volley forehand and ten volley backhand in front of the tennis net. Tennis balls were thrown from the right and the left side of the net, while standing parallel to the net, the player made a short movement with the racket in front of him/her, causing the ball to bounce and fall. The participant hit a ball which was caught by a special net. The forehand tennis stroke is made with the dominant hand. The racket was placed on

the dominant side; then, it was directed towards the ball. After the racket made contact with the ball, the racket was directed to the opposite arm of the player in a way of swinging. While performing a two-handed backhand stroke, the racket was held with a continental grip. It was placed on the opposite side to the dominant one. After the racket made contact with the ball, it was directed to the dominant side. In the case of a one-handed backhand, the racket was held with a dominant hand. These two types of strokes are presented in Figure 2. It is worth indicating that forehand and forehand volley are very similar moves in a certain part of the movement. The same goes for backhand strokes.

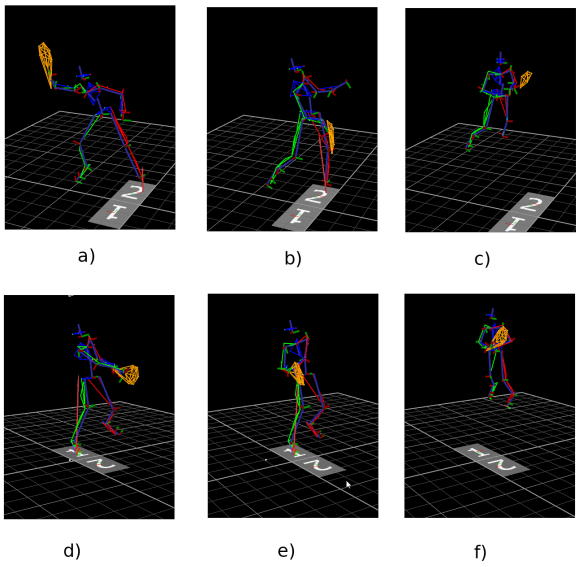


Figure 2. An example of forehand and backhand strokes (**a**) forehand preparation phase (**b**) forehand shot (**c**) no shot (**d**) backhand preparation phase (**e**) backhand shot (**f**) no shot.

Each performed stroke has been verified by a specialist. All failed strokes were rejected. Due to the fact that professional tennis players participated in the study, the well-performed strokes were repetitive.

2.3. Data Post-Processing

The Vicon Nexus software was used for post-processig of all obtained recordings. This tasks involved the following steps: marker labelling, gap filling using interpolation methods implemented in Vicon Nexus software (Pattern Fill and Rigid Body Fill), data cleaning, and applying the Plug-in-Gait model. The last one was only for the model representing human body. Additionally, a new model, consisting of all markers attached to the racket, was generated. The data prepared in this way was saved to c3d file.

The whole gathered recordings was verified by a professional tennis coach. As a result, the following number of tennis moves was obtained: backhand—212, forehand—197, forehand volley—180, backhand volley—180.

2.4. Attention Temporal Graph Convolutional Network

The idea of the A3T-GCN was taken from the work [31], where a similar structure was used to predict traffic volume in selected cities. The basic modification of this network consists of transforming the element responsible for the prediction into a classifier. Additionally, the elements responsible for the separation of spatial and temporal features have also been adapted. In the original approach, the Gated Recurrent Unit (GRU) network was applied. Due to extensive structure of the GRU network, inadequate to the problem, we are analyzing in our work RNN network, often also called BiRNN or Bidirectional RNN. It is schematically shown in Figure 3. Moreover, the original prediction was based on a

Context Vector, while in case of this study additional Multilayer Perceptron was added on the output of the classifier. The whole network structure used in this study is presented in Figure 4.

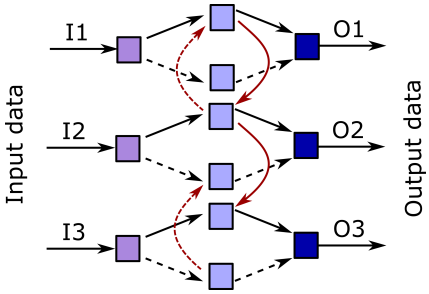


Figure 3. Scheme of the used BiRNN network.

The input data is arranged in a way of a graph $G = (V, E)$ consisting of M nodes. They denotes M joints and their position changes in time. The M value was equal to 39 for the study without a tennis racket and 46 in case of experiments with it. Each node is described as a set of three-dimensional values $V = \{v_{ti} | t = 1, ..., T, i = 1, ..., M\}$.

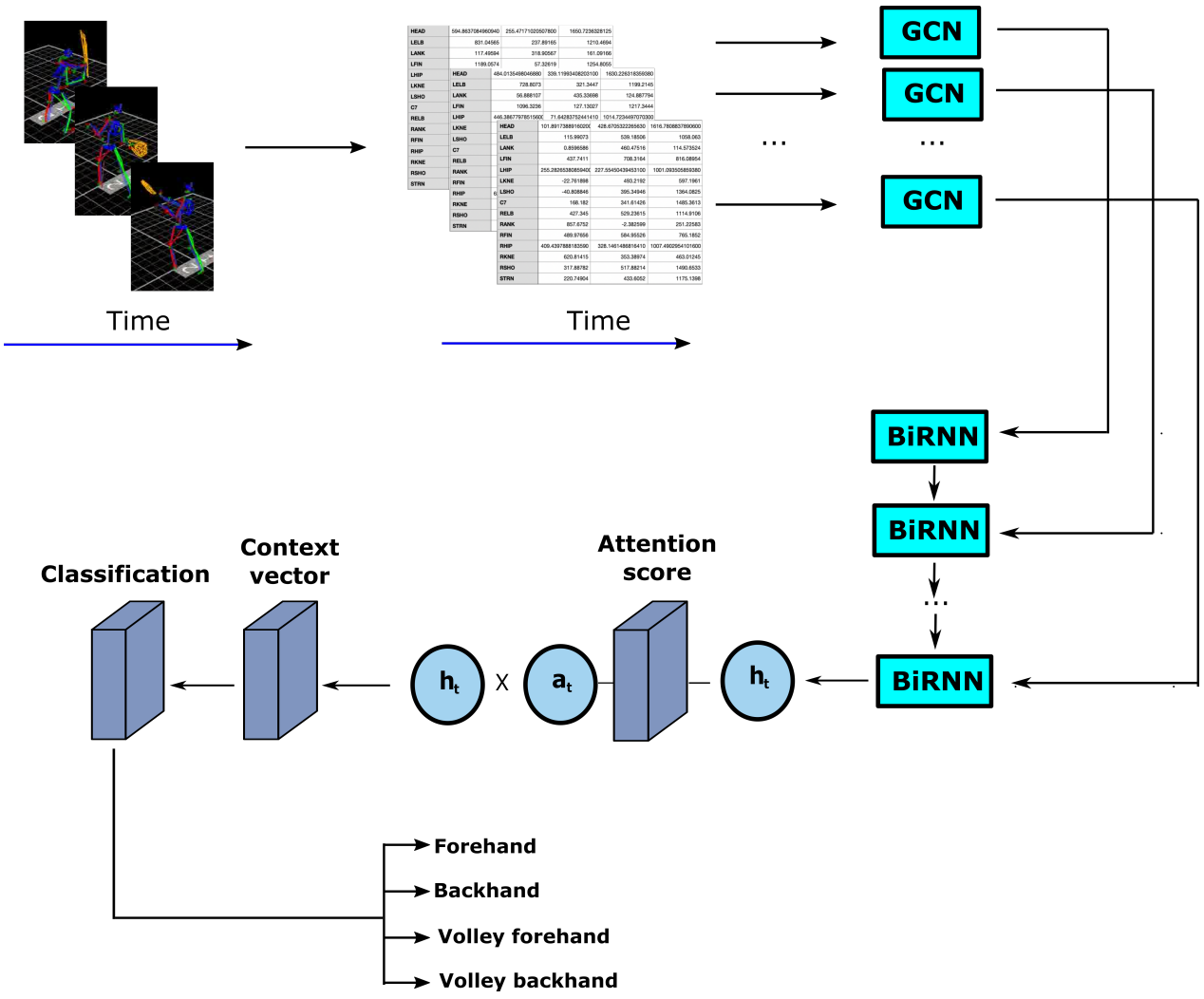


Figure 4. Classification model for tennis data movements.

2.4.1. Spatial Features

Usually, skeleton data studies are based on images or video as input, so the data are processed by typical Convolutional Neural Networks (CNN). In case of this study, as input data points in three-dimensional space were used, the proposed classifier was based on Graph Convolutional Networks (GCNs). The connections between the nodes of the G graph were presented in the form of the adjacency matrix A. The entire feature matrix has been marked with the X variable. To process graph nodes, the GCN network, uses a Fourier filter to determine the spatial relation between features. This relationship was characterized by Equation (1), which actually defines a multilayer GCN model.

$$F^{(n+1)} = \sigma\left(\widetilde{T}^{-\frac{1}{2}}\widetilde{O}\widetilde{T}^{-\frac{1}{2}}F^{(n)}\Theta^{(n)}\right) \tag{1}$$

where n represents the number of hidden layers, $\widetilde{O} = O + I_N$ is the adjacent matrix (O) with added self-connections, I_N describes the identity matrix, $\widetilde{T} = \sum_j \widetilde{O}_{ij}$, $F^{(n)}$ defines the output of n layer, $\Theta^{(n)}$ is a matrix which contains all parameters of specified n^{th} layer and $\sigma(\cdot)$ represents the sigmoidal function for a nonlinear model [32].

In this study, the GCN network consists of three layers. This structure can be described by Equation (2).

$$f(I,O) = \sigma\left(\widehat{O}ReLU\left(\left(\widehat{O}I\Psi_0\right)\widehat{O}I\Psi_1\right)\Psi_2\right) \tag{2}$$

where $\widehat{O} = \widetilde{T}^{-\frac{1}{2}}\widetilde{O}\widetilde{T}^{-\frac{1}{2}}$ indicates the preliminary step, $\Psi_0 \in R^{PxF}$ denotes the weight matrix between input and hidden layer, P defines the size of the feature matrix, while F is a value related to the number of the hidden unit, $\Psi_1, \Psi_2 \in R^{FxZ}$ define the weight matrices from hidden to output layer, $f(I,O) \in R^{NxZ}$, denotes the output length Z and $ReLU()$, is the Rectified Linear Unit, commonly used as neurons activation function [32].

2.4.2. Temporal Features

To indicate temporal features, which are the key elements in recognizing the analyzed types of tennis strokes, a BiDirectional Recurrent Neural Network was used. BiRNNs were applied to obtain the information about the player at time t. To gather this kind of data, the information about previous (in time $n-1$, $n-2$,...$n-n_f$, where n_f denotes the maximum number of frames in all c3d file) features were taken into consideration. If analyzed file had fewer frames the missing values were set to 0. The structure of whole temporal features elements can be expressed by Equations (3)–(6) [31]:

$$ugc_t = \sigma(W_u * [X_t, h_{t-1}]) \tag{3}$$

$$rgc_t = \sigma(W_r * [X_t, h_{t-1}]) \tag{4}$$

$$mc_t = \tanh(W_c[X_t, (rgc_t * h_{t-1})]) \tag{5}$$

$$h_t = ugc_t * h_{t-1} + (1 - ugc_t) * mc_t \tag{6}$$

where ugc_t denotes the update gate, which role is connected with controlling the information quantity at the previous moment, rgc_t indicates the reset gate, which is responsible for neglecting the state information at the previous moment, mc_t describes stored memory content at the current moment and h_t defines the output value at the current moment. W_u, W_r and W_c represent the weights in the training process for the updated gate layer, reset gate layer and output layer, respectively.

2.4.3. Attention Model

Commonly attention model is defined as an encoder–decoder. It is widely used in such applications as: traffic forecasting [31], image labeling [33], recommendation systems [34] or document classification [35]. Based on [36], it can be stated that the most general division of that kind of model includes hard and soft attention. In this study, the soft one was applied. The attention model's first application is to store information about the player's model. The second is to indicate the context vector, which is responsible for determining the predicted tennis player positions. The applied attention model consists of the following steps.

1. First, determine, using the BiRNN network, the successive hidden states $u_k(k = 1, ..., m)$ of the time series $I_k(k = 1, 2, ..., m)$, where m is the number of frames in series. As a result, the set of u_k states is defined.
2. Second, a context vector (C_v) is determined. In particular, the value of position change was determined on a basis of two hidden layers. Their features are indicated applying Equation (7):

$$e_i = \psi_{(2)}\left(\psi_{(1)}H + b_{(1)}\right) + b_{(2)} \tag{7}$$

where $\psi_{(1)}$ and $b_{(1)}$ denote the weight and bias of the first layer and $\psi_{(2)}$ and $b_{(2)}$ are similarly features of the second layer. H is a matrix with hidden layer values. To determine the values of $\psi_{(1)}, \psi_{(2)}$ the $Softmax$ function (8) is used.

$$Softmax = \frac{exp(e_i)}{\sum_{k=1}^{n} exp(e_k)} \tag{8}$$

Final, the C_v is defined as follows:

$$C_v = \sum_{i=1}^{n} Softmax * h_i \tag{9}$$

The final classification is performed by two-layer perceptron. This neural structure consists of one element in the first layer and four (related to four recognised strokes) in the second one. The $Softmax$ function is used to activate the neurons in the first layer, while the second layer is activated by a linear one.

2.5. Experiment

In this study, the forehand, backhand, volley forehand and volley backhand strokes were recognized. The whole tennis movements dataset consisted of backhand—212, forehand—197, volley forehand—180, and volley backhand—180. It represented the player's silhouette together with a tennis racket. Two types of experiments were performed. The first one concerned the whole set of data while the other only the player's silhouette by removing the coordinations of the tennis racket subject.

A series of experiments were carried out, taking into account the random division of data into the training, validation and test sets: 60%, 20% and 20%, respectively. The data was chosen from every type of stroke in the above-mentioned proportions. For each division, 20 tests were carried out, independently.

3. Results

Grouped results were presented in Tables 1–5. Selected parameters of the learning process showing the correctness of the model were shown in Figure 5. The loss value L_{CE} was calculated on a basis of the Sparse Categorical Cross-Entropy defined by Equation (10).

$$L_{CE} = -\sum_{i=1}^{n} T_i log(p_i), \tag{10}$$

for n classes, where T_i is a ground truth, p_i is the $Softmax$ probability for the ith class.

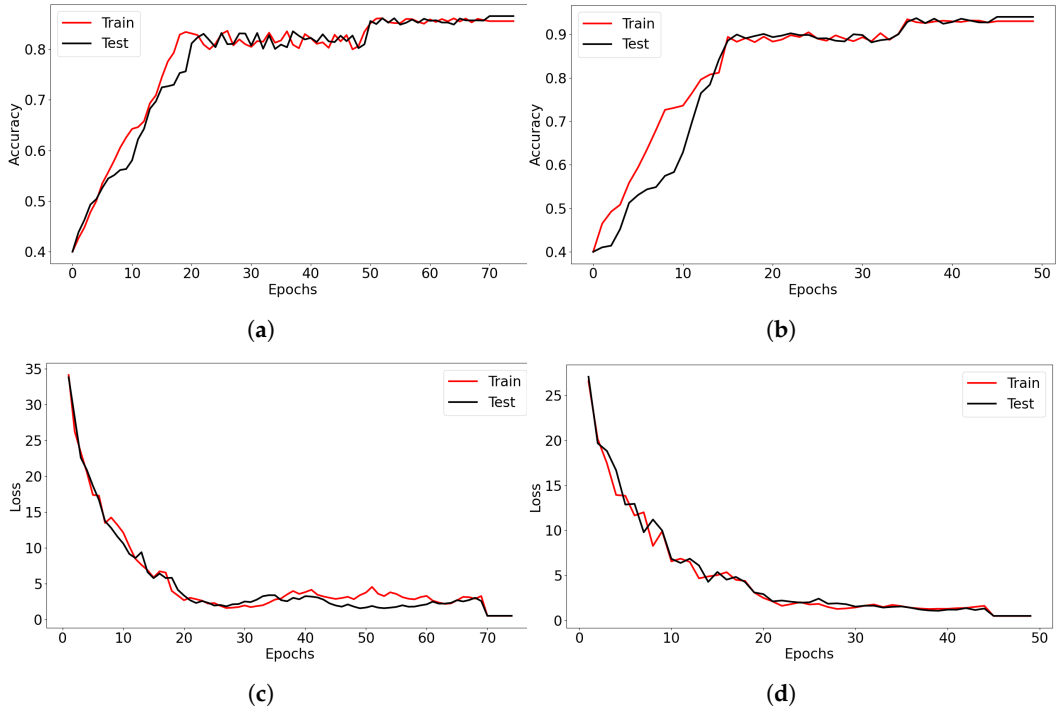


Figure 5. Selected learning parameters. (**a**) Learning accuracy for input without the tennis racket. (**b**) Learning accuracy for input with the tennis racket. (**c**) Loss value for input without the tennis racket. (**d**) Loss value for input with the tennis racket.

The assessment of the classifier quality was based on several standard measures, such as: Accuracy (11), Precision (12), Recall (13) and F1 score (14).

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ classifications} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{14}$$

where TP denotes the true positive fraction, FP—the false positive fraction, and FN—the false negative fraction.

Although Precision, Recall, and F1 are usually presented for binary classification, there is a simple way for extending their definition to multiple classes. In this case, Precision, e.g., for backhand, will be defined as correctly classified backhand strokes out of all classified backhand strokes. The Recall for backhand is the number of correctly predicted backhand strokes out of all input backhand strokes.

For the obtained accuracy results for two types of moves: strokes without and with a tennis racket (Table 1), the T-Test was calculated, for which $t = -8.2753$ was obtained, for $\alpha = 0.05$. The obtained result allowed us to state that it cannot be concluded that there is a difference between the means.

Table 1. Obtained Accuracy results for A3T-GCN.

Type of Input	Mean	Max	Min	±SD
Without racket	82.60%	85.54%	78.00%	2.08%
With racket	88.95%	93.00%	85.62%	2.62%

Table 2. Obtained Accuracies results for individual strokes.

Type of Input	Stroke	Mean	Max	Min	±SD
Without racket	Forehand	81.99%	85.54%	78.33%	2.36%
	Backhand	81.38%	85.25%	78.00%	2.49%
	Volley Forehand	82.72%	85.54%	78.26%	1.89%
	Volley Backhand	84.39%	85.43%	79.23%	2.02%
With racket	Forehand	89.84%	93.98%	85.89%	2.71%
	Backhand	88.41%	93.36%	85.71%	2.62%
	Volley Forehand	88.49%	93.87%	85.62%	2.82%
	Volley Backhand	89.08%	92.11%	86.60%	2.01%

Table 3. Obtained Precision results.

Type of Input	Stroke	Mean	Max	Min	±SD
Without racket	Forehand	86.99%	88.54%	82.98%	1.17%
	Backhand	80.99%	84.31%	76.41%	2.00%
	Volley Forehand	83.57%	85.85%	79.59%	2.00%
	Volley Backhand	84.50%	87.63%	78.00%	3.18%
With racket	Forehand	93.08%	97.89%	87.62%	4.38%
	Backhand	91.04%	94.90%	87.63%	2.34%
	Volley Forehand	89.45%	93.94%	85.00%	2.97%
	Volley Backhand	89.52%	93.00%	85.86%	2.28%

Table 4. Obtained Recall results.

Type of Input	Stroke	Mean	Max	Min	±SD
Without racket	Forehand	81.95%	85.00%	75.73%	2.95%
	Backhand	84.69%	87.76%	80.41%	2.37%
	Volley Forehand	87.22%	88.54%	83.87%	1.44%
	Volley Backhand	82.19%	85.00%	77.23%	2.22%
With racket	Forehand	89.09%	93.94%	85.00%	2.62%
	Backhand	89.71%	93.94%	85.86%	2.86%
	Volley Forehand	93.54%	97.89%	88.54%	3.96%
	Volley Backhand	90.76%	94.93%	86.73%	2.56%

Table 5. Obtained F1 results.

Type of Input	Stroke	Mean	Max	Min	±SD
Without racket	Forehand	84.39%	86.73%	79.19%	2.37%
	Backhand	82.77%	86.00%	78.39%	2.20%
	Volley Forehand	85.35%	87.18%	81.68%	1.17%
	Volley Backhand	83.33%	86.29%	77.61%	2.67%
With racket	Forehand	91.03%	95.88%	86.29%	3.38%
	Backhand	90.37%	94.42%	86.73%	2.58%
	Volley Forehand	91.44%	95.87%	86.73%	3.36%
	Volley Backhand	90.14%	93.48%	86.24%	2.40%

In order to check the correctness of the developed model, Leave-One-Out Cross-Validation (LOOCV) was performed (Table 6). This is a computationally expensive procedure; however, it allows us to obtain clear and unbiased information about the model. Using LOOCV, the root mean squared error (RMSE) for n tests was determined:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{15}$$

where n—denotes number of test, y_i—true value, $\widehat{y_i}$—predicted value.

$$RMSE = \sqrt{MSE} \tag{16}$$

Table 6. Obtained Values for LOOCV.

Value	Input without Racket	Input with Racket
RMSE	9.64%	5.31%
SD RMSE	±5.58%	±5.79%

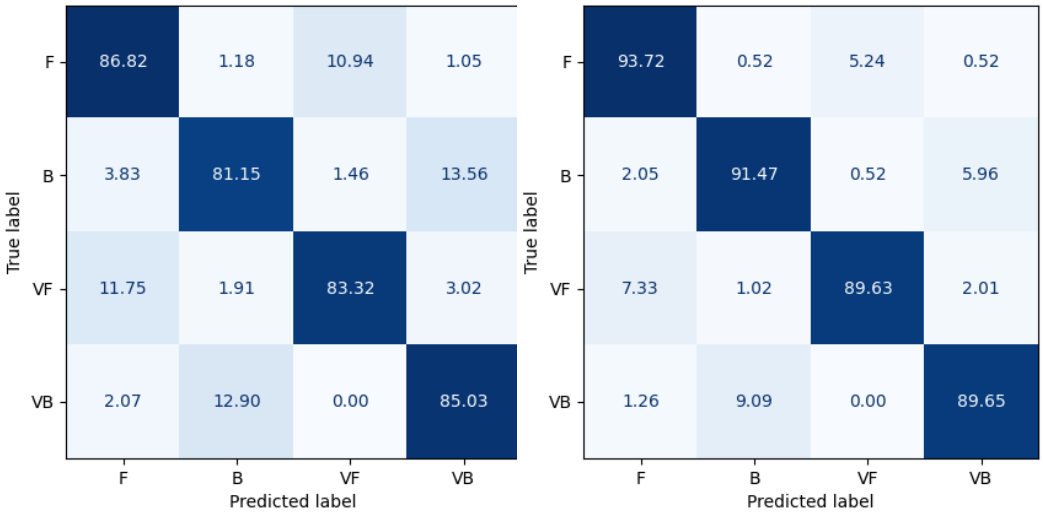
4. Discussion

In this study, the recognition of tennis forehand, backhand, volley forehand and volley backhand was performed based on data gathered in c3d files in a form of three-dimensional coordinates. Both the player's silhouette indicated by 39 markers and the silhouette together with a tennis racket represented by 7 additional markers were analyzed.

As it can be observed in Tables 1 and 2 and Figure 6 the mean accuracy depends on the captured data. It is higher for the experiments with the whole tennis player's silhouette together with a tennis racket than the experiment involving only a single body. It can be concluded that the arrangement and trajectory of a tennis racket plays an extremely important role in the correct classification.

Furthermore, in the case of Precision (see Table 3), the obtained mean results are higher for the combination of the tennis player's body model with a tennis racket for all analyzed strokes. The same dependence applies to Recall results (Table 4) and the F1 score (Table 5).

It should be noted that the standard deviation, amounting to a few percent, is low for all the obtained results, which shows the stability of the applied classifier.



(**a**) (**b**)

Figure 6. Confusion matrices (in %). (**a**) Matrix for input without the tennis racket. (**b**) Matrix for input with the tennis racket.

In Table 7, the state-of-the-art studies related to the tennis strokes recognition are presented. They were performed using various types of data obtained from different sources, such as sensors, video or motion capture systems. The most research in this field were carried out on the well-known THETIS database. Both the data in the form of video, as well as the images obtained from the Kinect motion capture system, were the source for recognizing tennis movements. Broadcast video involving real matches or tournaments with top tennis players was also often taken into consideration. Various types of neural network approaches were used for these purposes. It is worth stressing that graph neural networks (ST-GCN and A3T-GCN) were used to recognize basic tennis strokes based on data obtained from an optical motion capture system. This classifier was chosen due to the characteristics of the recorded data. The applied human model, represented by 39 markers attached to the body in fixed locations, is transformed into a graph, which reflects the topology of the human silhouette. This approach allowed us to obtain a high accuracy. Analyzing the results presented in the study [28], it can be seen that used the A3T-GCN classifier allows for better recognition of tennis strokes than the ST-GCN one, despite the different types of input data. In the previous study, described in [28], the accuracy was obtained at the level of 68.9% for the 60% of data belonging to the training set, which corresponds to the settings in this study. The forehand and backhand classification was performed using images containing the subjects of the tennis player together with the racket. They were generated based on three-dimensional data by Vicon Nexus software. Based on this kind of input data the simplified model, both for tennis player and a racket, was created. The tennis racket was represented only by two points referring to its head and handle. The mean accuracy obtained in this study is higher for both the analyzed silhouette and its combination with a tennis racket compared to the results obtained in the work [28]. In [30], the study concerned the images obtained from three-dimensional data were analysed. The classification of forehand, divided into preparation and the hit phases, and backhand, also divided into preparation and the hit phases, as well as no-hit was performed using the Attention Temporal Graph Convolutional Network. The achieved accuracy results in a form of mean of two phases for forehand stroke did not exceed 80% while for backhand—77%. The obtained mean accuracy results in this paper are higher for the analyzed strokes with and without a tennis racket. The studies presented in this paper concerning tennis movements recognition was performed based on three-dimensional data in the form of coordinates of markers placed on the player's body and a tennis racket was used. As it can be seen in Table 7 this approach is unique.

Table 7. Results comparison with the state-of-the-art. FH—forehand, BH—backhand, S—serve, BS—backspin, SM—smash, V—volley, H—hit, NH—no hit.

Data/Dataset	Type of Input	Classified Types of Tennis Movements	Detection Method	Accuracy	Paper
SensorTile	signal	FH BH, S	DNN	94–97%	[14]
SensorTile	signal	FH, BH, FH, BH	SVM NN DT RF kNN	90.82–98.86% 98.76–100% 84.69–95.54% 93.75–98.96% 87.76–99.44%	[15]
IMU	sensor	FH, BH, BS, S, SM	Pan Tompkins algorithm	80.6–98.1%	[16]
THETIS	video	BH, V, FH, V, S, SM,	LSTM	81.23–89.42%	[19]
THETIS	video	BH FH, V, S, SM	SVM CRF	51.20% 86.44%	[17]
THETIS HMDB51	video	BH, V FH, V S, SM	Deep Historical LSTM	62% 54%	[21]
THETIS KTH	video	BH V, S, SM	LSTM	70.17–97.67%	[22]
THETIS KTH	video	BH FH, V, S, SM	SVM	53.08–60.23% 90.65%	[18]
KTH	Video	S, H, NH	KLDA	73.34–92.29%	[23]
Broadcast	video	F, B	SVM	90.21%	[24]
Broadcast	video	F, B	SVM	87.10%	[25,26]
mixed	signal video	F, B, S	SVM, kNN SVM kNN kNN	89.69–97.02% 95–98.67% 82.43–88.36% 89.41–93.44% 84.73–100%	[27]
Vicon with fuzzy input	images	FH, BH, NH	ST-GCN	86.3–87.3%	[28]
Vicon	images	F, B, NH	ST-GCN	64.1–74.3%	[28]
Vicon with fuzzy input Vicon	images images	F, B, NH F, B, NH	A3T-GCN A3T-GCN	86.9–93.82% 74.22–81.95%	[30]
Vicon Vicon	**3d silhouette** **3d silhouette & racket**	F, B, V F, B, V	A3T-GCN A3T-GCN	**78.33–85.54** % **85.62–93.98**%	**Our work**

The results obtained in this paper suggest that the type of input data affects the accuracy of tennis stroke classification. Whole body data stored in the form of three-dimensional coordinates allows to achieve better results than in the case of images obtained

from three-dimensional data. In addition, the inclusion of a tennis racket in the input data improves the classification quality of these strokes.

5. Conclusions

The state-of-the-art of this study was to verify the impact of adding a tennis racket to the input data of the whole player's silhouette on the final classification of four main tennis strokes. For the purpose of this study the A3T-GCN classifier was applied. The tennis moves were represented in a form of three-dimensional motion data. The described approach gave satisfactory results. Forehand, backhand, volley forehand and volley backhand were taken into consideration. According to the previous authors' study considering the ST-GCN network as a classifier [28], the obtained accuracy for two basic tennis movements (forehand and backhand) recognition in this research has been improved. As the results showed, adding a tennis racket to the data presenting the whole body silhouette significantly improved the classification quality. While comparing two graph networks, i.e., A3T-GCN and ST-GCN, the obtained results in this study clearly indicated that the A3T-GCN structure might be considered as the most suitable for motion data expressed in a form of three-dimensional coordinations. Further work will focus on the possibilities of using other methods of computational intelligence, in particular deep learning methods, and investigating their impact on the efficiency of movement classification in sport. Moreover, research on the possibility of using aggregation of classification methods, may be of interest in further studies in this field.

Author Contributions: Conceptualization, M.S.-P. and P.P.; methodology, M.S.-P.; software, P.P.; validation, M.S.-P. and P.P.; formal analysis, M.S.-P.; investigation, P.P.; resources, M.S.-P. and P.P.; data curation, M.S.-P. and P.P.; writing—original draft preparation, M.S.-P. and P.P.; writing—review and editing, M.S.-P. and P.P.; visualization, M.S.-P.; supervision, M.S.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The research programme titled "Biomechanical parameters of athletes in the individual exercises" based on the analysis of 3D motion data and EMG, realised in the Laboratory of Motion Analysis and Interface Ergonomics was approved by the Commission for Research Ethics, No. 2/2016 dated 8 April 2016.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Student Sports Club Tennis Academy POLSART and Sport Club KS-WiNNER for their support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE PAMI* **2022**. [CrossRef] [PubMed]
2. Aggarwal, J.K.; Xia, L. Human activity recognition from 3d data: A review. *Pattern Recognit. Lett.* **2014**, *48*, 70–80. [CrossRef]
3. Host, K.; Ivašić-Kos, M. An overview of Human Action Recognition in sports based on Computer Vision. *Heliyon* **2022**, *2022*, e09633. [CrossRef]
4. Ma, C.; Fan, J.; Yao, J.; Zhang, T. NPU RGBD Dataset and a Feature-Enhanced LSTM-DGCN Method for Action Recognition of Basketball Players. *Appl. Sci.* **2021**, *11*, 4426. [CrossRef]
5. Guo, J.; Liu, H.; Li, X.; Xu, D.; Zhang, Y. An Attention Enhanced Spatial–Temporal Graph Convolutional LSTM Network for Action Recognition in Karate. *Appl. Sci.* **2021**, *11*, 8641. [CrossRef]
6. Qi, M.; Wang, Y.; Li, A.; Luo, J. Sports Video Captioning via Attentive Motion Representation and Group Relationship Modeling. *IEEE Trans. Circuits Syst. Video Technol. IEEE Trans. Circ. Syst. Vid.* **2020**, *30*, 2617–2633. [CrossRef]
7. Martinez, B.; Modolo, D.; Xiong, Y.; Tighe, J. Action recognition with spatial-temporal discriminative filter banks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5482–5491.

8. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
9. Chen, X.; Pang, A.; Yang, W.; Ma, Y.; Xu, L.; Yu, J. SportsCap: Monocular 3D Human Motion Capture and Fine-grained Understanding in Challenging Sports Videos. *IJCV* **2021**, *129*, 2846–2864. [CrossRef]
10. Nan, M.; Trăscău, M.; Florea, A.M.; Iacob, C.C. Comparison between Recurrent Networks and Temporal Convolutional Networks Approaches for Skeleton-Based Action Recognition. *Sensors* **2021** *21*, 2051. [CrossRef]
11. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
12. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
13. Liu, J.; Che, Y. Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network. *J. Electron. Imaging* **2021**, *30*, 33017. [CrossRef]
14. Ganser, A.; Hollaus, B.; Stabinger, S. Classification of Tennis Shots with a Neural Network Approach. *Sensors* **2021**, *21*, 5703. [CrossRef] [PubMed]
15. Ma, K. A Real Time Artificial Intelligent System for Tennis Swing Classification. In Proceedings of the IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Herlany, Slovakia, 21–23 January 2021; pp. 21–26.
16. Pei, W.; Wang, J.; Xu, X.; Wu, Z.; Du, X. An Embedded 6-axis Sensor based Recognition for Tennis Stroke. In Proceedings of the IEEE International Conference on Consumer Electronics, ICCE 2017, Taipei, Taiwan, 29 March 2017; pp. 55–58.
17. Vainstein, J.; Manera, J.; Negri, P.; Delrieux, C.; Maguitman, A. Modeling video activity with dynamic phrases and its application to action recognition in tennis videos. In Proceedings of the Iberoamerican Congress on Pattern Recognition, Puerto Vallarta, Mexico, 2–5 November 2014;; Springer: Berlin/Heidelberg, Germany, 2014; pp. 909–916.
18. Gourgari, S.; Goudelis, G.; Karpouzis, K.; Kollias, S. Thetis: Three dimensional tennis shots a human action dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, ON, USA, 23–28 June 2013; pp. 676–681.
19. Mora, S.V.; Knottenbelt, W.J. Deep learning for domain-specific action recognition in tennis. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 170–178.
20. Mora, S. *Computer Vision and Machine Learning for In-Play Tennis Analysis: Framework, Algorithms and Implementation*; University of London, Imperial College of Science, Technology and Medicine, Department of Computing: London, UK, October 2017.
21. Cai, J.; Hu, J.; Tang, X.; Hung, T.-Y.; Tan, Y.-P. Deep Historical Long Short-Term Memorys for Action Recognition. *Neurocomputing* **2020**, *407*, 428–438. [CrossRef]
22. Ullah, M.; Mudassar Yamin, M.; Mohammed, A.; Daud Khan, S.; Ullah, H.; Alaya Cheikh, F. Attention-based LSTM network for action recognition in sports. *Electron. Imaging* **2021**, *6*, 302-1–302-5. [CrossRef]
23. Faraji Davar, N.; De Campos, T.; Kittler, J.; Yan, F. Transductive transfer learning for action recognition in tennis games. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 7 November 2011; pp. 1548–1553.
24. Zhu, G.; Xu, C.; Huang, Q.; Gao, W.; Xing, L. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In Proceedings of the 14th ACM international conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 431–440.
25. Zhu, G.; Xu, C.; Gao, W.; Huang, Q. Action recognition in broadcast tennis video using optical flow and support vector machine. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 13 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 89–98.
26. Zhu, G.; Xu, C.; Huang, Q.; Gao, W. Action recognition in broadcast tennis video. In Proceedings of the International Conference on Pattern Recognition, Hong Kong, China, 18–24 August 2006; pp. 251–254.
27. Conaire, C.Ó.; Connaghan, D.; Kelly, P.; O'Connor, N.E.; Gaffney, M.; Buckley, J. Combining inertial and visual sensing for human action recognition in tennis. In Proceedings of the first ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, Firenze, Italy, 29 October 2010; pp. 51–56.
28. Skublewska-Paszkowska, M.; Powroznik, P.; Lukasik, E. Learning three dimensional tennis shots using graph convolutional networks. *Sensors* **2020**, *20*, 6094. [CrossRef]
29. Skublewska-Paszkowska, M.; Powroznik, P.; Karczmarek, P.; Lukasik, E. Aggregation of Tennis Groundstrokes on the Basis of the Choquet Integral and Its Generalizations. In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 14 September 2022; pp. 1–8.
30. Skublewska-Paszkowska, M.; Powroznik, P.; Lukasik, E. Attention Temporal Graph Convolutional Network for Tennis Groundstrokes Phases Classification. In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 18–23 July 2022; pp. 1–8.
31. Bai, J.; Zhu, Y.; Song, Y.; Zhao, L.; Hou, Z.; Du, R.; Li, H. A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting. *ISPRS Int. J.—Geo-Inf.* **2021**, *10*, 485. [CrossRef]

32. Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Deng, M.; Li, H. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 3838–3858. [CrossRef]
33. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, R.; Show, Y. Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
34. Xiao, J.L.; Ye, H.; He, X.; Zhang, H.; Wu, F.; Chua, T. Attentional Factorization Machines: Learning the Weight of Feature, Interactions via Attention Networks. *arXiv* **2017**, arXiv:1708.04617.
35. Pappas, N.; Popescu-Belis, A. Multilingual Hierarchical Attention Networks for Document Classification. *arXiv* **2017**, arXiv:1707.00896.
36. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.