International Journal of Computer Science in Sport

Volume 12/2013/Edition 1

TABLE OF CONTENTS

Editorial

Arnold Baca

Department of Biomechanics, Kinesiology and Applied Computer Science,

ZSU, University of Vienna

Dear readers:

Welcome to the summer 2013 issue of the **International Journal of Computer Science in Sport (IJCSS).**

The current issue is subdivided into 2 parts including six original papers.

The first part contains two research papers and one project report.

Nicolas Houel, Antoine Faury and **Didier Seyfried** illustrate a comparative study regarding the influence of the point of attachment of two different sensors on the assessment of squat jump performances.

Stephen J. Robinson presents probalibilistic optimization routines for finding a perfect lineup having the highest number of runs per game in youth baseball.

The project report by **Jürgen Perl, Andreas Grunz** and **Daniel Memmert** introduces a novel approach for evaluating the tactical performance in soccer. The method combines pattern-based tactics analysis with success-oriented statistical frequency analysis.

Three additional articles are appended thereafter in the second part ("Special Edition").

In this "Special Edition" part of the IJCSS selected scientific papers presented at the 9th Symposium of the Section Computer Science in Sport (Sportinformatik) of the German Association of Sport Science (dvs) are included. The contributions underwent a further review process to ensure the quality of the scientific content. IACSS would like to thank the entire committee under the guidance of Prof. Dietmar Saupe for the organization of the conference and the support in the selection process of the papers.

The investigation made by **Oliver Hummel, Ulrich Fehr** and **Katja Ferger** demonstrates the feasibility and potential of smartphones as a low-cost alternative for performance diagnostics.

Heike Leutheuser and **Bjoern M. Eskofier** describe their examination on how different heart rate variability (HRV) parameters change during one hour of running involving datasets of 295 athletes.

The study by **Josef Wiemeyer** examines the influence of previous game experience and the presence of music and sound on game performance and game experience.

If you have any questions, comments, suggestions and points of criticism, please send them to me.

Enjoy the summer!

Arnold Baca, Editor in Chief
University of Vienna, arnold.baca@univie.ac.at

PART 1

APPENDED PAPERS

Influence of the Point of Attachment of two Accelerometers on the Assessment of Squat Jump Performances

Nicolas Houel, Antoine Faury & Didier Seyfried

Research Department, French National Institute of Sport, Expertise and Performance (INSEP), Paris, France

Abstract

The aims of this study were to compare the validity of two accelerometers with a force plate and to determine the influence of the sensor's point of attachment on the assessment of squat jump performances. Nine male subjects performed a number of squat jumps (n = 38) on a force plate, either with a Myotest or Mensense system attached to their hips and backs. For evaluation purposes, two-way ANOVA tests, correlation coefficients and Bland and Altman tests were used to compare the influence between each sensor and the force plate based on the flight time as well as maximal and take off velocities. The obtained results showed that each sensor, the point of attachment on the subject and their interaction significantly influenced the assessment and validity of the flight time as well as maximal and take off velocities. When attached to the subject's back, the Mensense sensor estimated the flight time with the best validity in comparison to the force plate measurement. On the other hand, the Myotest sensor estimated the maximal velocity with the best validity when it was attached to the subject's hip. The take off velocity was estimated by both sensors with very low accuracy. It was therefore inferred that the point of attachment of the sensors and the computational software have a direct influence on the assessment of squat jump performances.

KEYWORDS: MYOTEST, ACCELEROMETER, SQUAT JUMP, VALIDITY

Introduction

Vertical jumps are widely used for testing the joint's coordination (Bobbert et al., 2008; Nuzzo et al., 2011), physical activity and training level of sportsmen (Quagliarella et al., 2006; Dionyssiotis et al., 2009). Flight time (Nuzzo et al., 2008; Glatthorn et al., 2011; Castagna et al., 2012), maximal and take off velocities (Dionyssiotis et al., 2009; Casartelli et al., 2010; Houel et al., 2010 et al., 2011) and jump height (Dowling & Vamos, 1993; Hatze, 1998; Kibele, 1998; Garcia-Lopez et al., 2005; Picerno et al., 2011) are commonly taken as significant variables for the assessment of squat jump performances. Various devices and methodologies are applied to determine such parameters (Nuzzo et al., 2008; Crewther et al., 2011; Castagna et al., 2012). Among them, the application of force plates is generally considered as one of the best methods for the assessment of these determinants (Hatze, 1998; Kibele, 1998; Bobbert et al., 2008; Nuzzo et al., 2008; Samozino et al., 2008; Crewther et al., 2011; Casartelli et al., 2010). However, this kind of equipment requires performing the measurement in laboratory or under very constraining conditions, which are not always in agreement with sport training or mobile applications (Quagliarella et al., 2006; Crewther et al., 2011; Glatthorn et al., 2011).

There are also a variety of portable instruments for the estimation of vertical jump performances. For example, contact mats and optical cells are used to estimate vertical jump performances in sport training (Castagna et al., 2012, Garcia-Lopez et al., 2005, Glatthorn etal., 2011, Hatze, 1998). The flight time values gathered from contact mats show a strong correlation with those from force plates. At the same time, contact mats appear to have different flight time values compared with force plates (Garcia-Lopez et al., 2005). In addition, they are not recommended for evaluating centre of mass variables of subjects during single vertical jumps (Hatze, 1998). Also optical cells show strong correlations estimating the vertical jump height and flight time in comparison with force plates (Glatthorn et al., 2011; Castagna et al., 2012). Systems like Optojump, though, demonstrate a lower jump height as well as flight time compared with force plates (Glatthorn et al., 2011; Castagna et al., 2012). Optical cells and contact mats are not to be used interchangeably with force plates when estimating single vertical jump performances and interpreting data of different devices (Garcia-Lopez et al. 2005; Glatthorn et al., 2011; Castagna et al., 2012). One reason for this could be explained by the assumption that the subject's take off and landing configurations are identical (Bosco et al., 1983). However, this assumption rarely occurs (Hatze, 1998; Garcia-Lopez et al., 2005, Glatthorn et al., 2011).

During the last ten years, small and convenient accelerometers have been increasingly used to measure human motion in sports (Jidovtseff et al., 2008; Innocenti et al., 2006; Houel et al., 2010; Crewther et al., 2011; Houel et al., 2011; Picerno et al., 2011; Castagna et al., 2012). Systems like Myotest SA are applied to measure specific human motions in sport training (e.g. squat jump, bench press etc.) using a triaxial accelerometer (Jidovtseff et al., 2008; Crewther et al., 2011; Houel et al., 2011). The study by Crewther et al. (2011) focuses on the accuracy measurement of these devices during squat press motions. The accelerometer is fixed to a bar with different loads. The squat press motion is performed by sportsmen with the bar attached directly to the force plate. The obtained results show systematic bias and relatively large random errors when assessing the accelerometer's peak forces for low loads. The authors discuss that the load could influence the bar's velocity and could change the estimation of the centre of mass of the subject (human and bar) on the force plate, which is not measurable by an accelerometer fixed to the bar. In the presented method, the processing of the accelerometer signals considers only human movements with respect to the bar and therefore may neglect key parameters. Other studies compare the accuracy and reliability of the Myotest Pro system attached directly to the subject during squat jumps (Casartelli et al., 2010; Castagna et al., 2012). The gathered results indicate that only the flight time values estimated by the Myotest Pro show large correlations with the Optojump system and a force plate. However, significantly low and systematic bias is observed for the flight time values between the Myotest Pro and Optojump systems or a force plate (Casartelli et al., 2010; Houel et al., 2011; Castagna et al., 2012).

The flight time as well as maximal and take off velocities can be estimated with different level of accuracy with other commercial accelerometers (Houel et al., 2010). In general, the control of the validity and reliability of these devices is a fundamental step for accepting them as appropriate systems for the assessment of kinetic data during vertical jumps. In particular, differences regarding the assessment of the accuracy and reliability of the flight time and velocity variables may occur between the used accelerometer devices. Such differences could be explained by the point of attachment of the sensor on the subject during the vertical jump motion. In addition, the sensor's orientation has a direct impact on the measurement accuracy. When the sensor is attached to the trunk, the accuracy of the jump height estimation using inertial sensors can be improved if the measured acceleration is corrected for the trunk rotations (Innocenti et al., 2006; Picerno et al., 2011). Picerno et al. (2011) investigate this method to estimate the jump height and flight time. Since the authors compare the sensor's jump height values with a stereophotogrammetric method, where light reflective markers are placed directly on the sensor, no relation between the jump height assessment of the sensor and the subject's centre of mass performance is explored. The flight time of the centre of mass can be accurately estimated using high speed cameras or sensors compared with force plates (Garcia-Lopez et al., 2005; Picerno et al., 2011). Only little research can be found regarding the assessment of the centre of mass velocity on the basis of simple and portable devices. Procedures using vertical velocity at take off seem to be more efficient for jump height assessment (Hatze, 1998; Kibele, 1998). One difficulty could be the limitation to asses and compute similar velocity and flight time values from the force plate using only a single sensor. As an example, force plates evaluate the total human body motion with respect to the centre of mass method (Hatze, 2008; Samozino et al., 2008). Accelerometers measure only the motion at the point of attachment. Thus, it is not possible to estimate the motion of the human's centre of mass, except if the point of attachment has similar motion characteristics and if special signal processing routines are applied. In the case of squat jumps, the accelerometers should be preferably attached to the hip and back bones since these segments might be the closest to the centre of mass. At the same time, the hip and back bones have the advantage to be close to the skin.

The initial research aim of this study was to define the influence of the sensor's point of attachment on the measurement accuracy. The second goal was to compare the validity of two commercial accelerometers commonly used in sport training (Myotest pro and Mensense) with the results obtained from a force plate. The measured flight time as well as maximal and take off velocities were compared in order to propose recommendations in the use of each device in respect to gait measurement standards and the equipment's point of attachment.

Methods

Nine male subjects with mean (± standard deviation SD) height of 179.8 ± 5.2 cm and body mass of 76.02 ± 6.9 kg voluntarily participated in this study. The participants were physical education students. Each subject was informed about the procedure of the study and signed a written informed consent. Before the test session, each subject performed a standard warm up program consisting of a 5 minutes run at their preferred submaximal velocity, 5 minutes submaximal squat jump familiarization on the force plate with the Myotest Pro (MYO) and Mensense (MEN) sensors and, finally, stretching the lower extremity muscles. For each squat jump, the beginning of the subject's motion was controlled with a bar placed under the legs with the aim to disallow a counter movement motion (Figure 1). The bar height was estimated for each subject before the first squat jump. The bar height was defined as the height where the subject initiates his squat jump impulse with a 90 degree knee angle. The knee angle was estimated using a Myotest goniometer (Acceltec SA). The subjects were asked to perform three squat jumps with both sensors (MYO and MEN) attached to their hips, and three additional squat jumps with both sensors attached to their backs. Thereby, the subjects were asked to keep their hands on their hips. The participants had a rest period of 3 minutes between each jump to limit the fatigue effect. The squat jumps were investigated with respect to the total impulses and landing phases. Jumps were selected only if they respect two conditions: i) only the feet are allowed to touch the force plate, ii) a counter movement is not performed during the impulse phase. A range of two to three jumps were selected for each subject. At the end, nineteen jumps were selected for each point of configuration. In sum, thirty-eight jumps were studied.

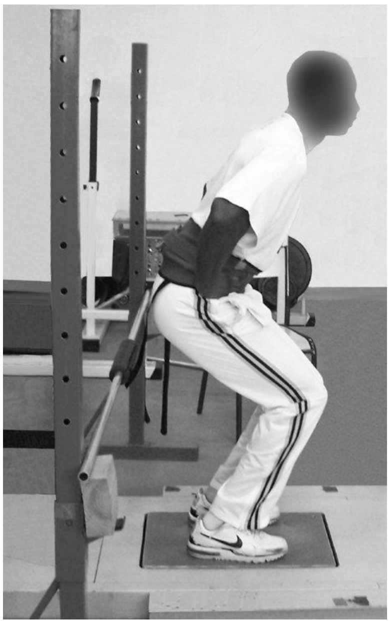

Figure 1. Experimental design.

All squat jumps were performed on a force plate (Kistler type 9281B). The sample frequency of the force plate was 500 Hz. In order to compute the acceleration for each squat jump, the gravitational force was divided by the body mass and subtracted from the vertical component of the force. The flight time was estimated on the basis of the acceleration data, where the acceleration values were nearly constant and equal or lower than - 9.81m.s^{-2} (Figure 2). The velocity data was computed using numerical integration (trapezoidal rule) of the vertical acceleration data during the impulse phase of the squat jump (Dowling & Vamos, 1993; Kibele, 1998). The subject's maximal velocity (Vmax in m.s^{-1}) was estimated at the point where the acceleration data was equal to zero at the end of the impulse phase. The subject's take off velocity (Vtoff, in m.s^{-1}) was estimated at the moment when the acceleration data was almost equal to the constant g = -9.81 m.s^{-2} at the end of the impulse phase (before take off).

The MYO system includes a triaxial accelerometer sensor as well as a software package developed by Myotest SA. It allows the recording of the vertical acceleration at a sampling frequency of 500 Hz. For the first three jumps, the light weight (58 grams) MYO sensor was fixed vertically to the side of the athlete's hip on the left coxo-femoral joint in agreement with the Myotest SA recommendations. For the other jumps, the sensor was vertically attached to the athlete's back at the level of the L5 lumbar. During the training mode, the Myotest Pro software was able to automatically compute the vertical acceleration, force and velocity curves as well as the maximal velocity (Vmax). The take off point was estimated during the impulse phase on the basis of the acceleration data, where the acceleration was equal or closer to - 9.75 m.s^{-2} after the acceleration curve decreased under zero. The lift off was estimated during the landing phase on the basis of the acceleration data when the acceleration was equal or closer to - 9.81 m.s^{-2} before the acceleration curve increased to zero. The flight time was calculated by subtracting the lift off minus the take off values. The take off velocity data was computed at the point of take off using the velocity data of the Myotest software.

The MEN NanoIMU system (Mensense, USA) was linked to the MYO sensor. The MEN device includes a 20 gram triaxial accelerometer sensor, which allows three dimensional acceleration measurements at a sampling frequency of 150 Hz. The triaxial acceleration data was transmitted via USB to a portable computer. A software solution was developed in Matlab R2008a in order to diagnose the motion starting before the squat jump impulse using an acceleration threshold below 1.5 $m.s^{-2}$. The sensor orientation was computed during a period of 0.26 s when the subject was in the squat posture just before the impulse phase using rotation matrices. The vertical acceleration was estimated on the basis of the acceleration data and the sensor orientation. The signals were filtered using a second order Butterworth filter with 5 degree and 25 Hz cut-off frequency. In addition, a 500 Hz interpolation was applied on the vertical acceleration data using inverse Fourier transformation. The flight time as well as maximal and take off velocities were computed by the same processing method as for the force plate.

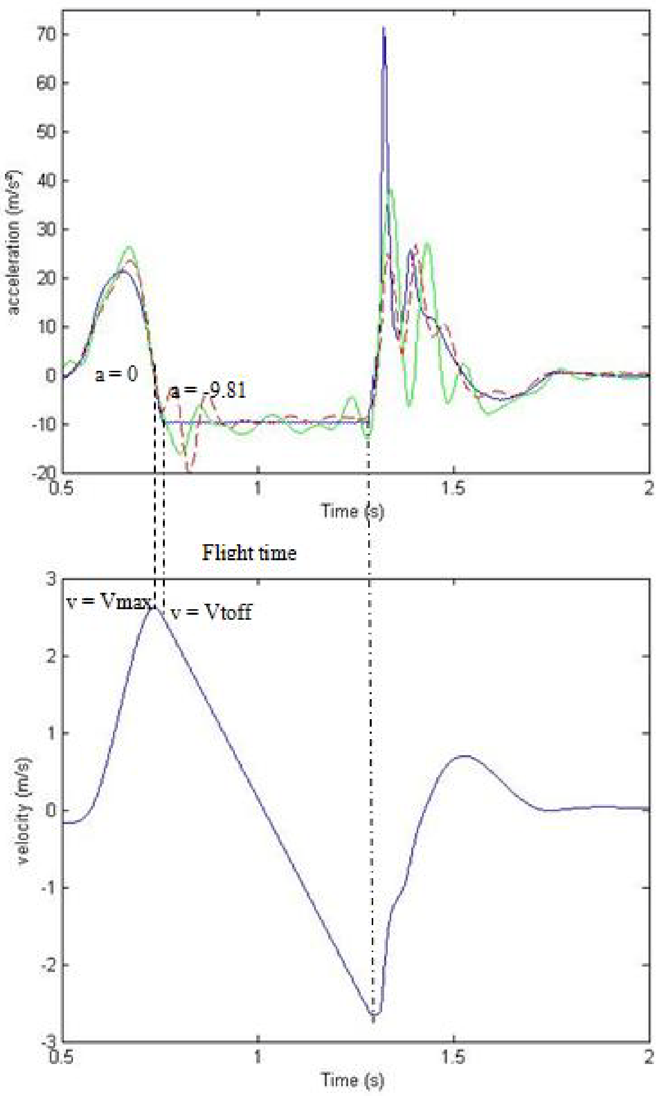

Figure 2. Acceleration (top) and velocity (bottom) data analysis during a squat jump when the sensors (MYO and MEN) were attached to the hip. The force plate data is illustrated in blue, the MYO data in red and the MEN data in green. The dotted black lines represent the relations between the acceleration and vertical data used to estimate Vmax, Vtoff and flight time.

The vertical acceleration data of the force plate (FP) and each sensor (MYO and MEN) were synchronized at the instant of Vmax – i.e. when the acceleration data was equal to zero before the subject took off (Figure 2).

A two-way ANOVA test was used to compare the significant differences between both devices (MYO and MEN) when attached either to the hip (H) or back (B). The FP was utilized to estimate the relevant kinetic variables (Vmax, Vtoff, t). Correlation coefficients (r) were used to compare the relation between the FP values and the data acquired from the sensors (MYO and MEN) and to estimate the flight time and velocity variables (Atkinson & Nevill, 1998). A Bland and Altman test was applied to define the accuracy and reliability between the FP and each sensor (Bland & Altman, 1995). A Lilliefors test was used to confirm the normality of the data (Lilliefors, 1967). No heteroscedasticity was observed when plotting the absolute differences against the individual means and calculating the correlation coefficients (Atkinson & Nevill, 1998).

Results

The obtained results showed that the flight time as well as maximal and take off velocities were estimated with different accuracy and reliability depending on the sensor and its point of attachment.

The two-way ANOVA test revealed that the MEN equipment was more influenced by the point of attachment to the subject and its interaction than the MYO device when considering kinetic variables. The point of attachment to the subject using the MEN system had, in comparison to the FP measurements, a significant effect on the estimation of Vmax ($p = 2.18*10^{-6}$), Vtoff ($p = 6.73*10^{-7}$) and t ($p = 0.0004$). On the other hand, the point of attachment to the subject using the MYO device had, in comparison to FP measurements, no significant effect on the estimation of Vmax ($p = 0.58$). The point of attachment to the subject using the MYO system had, in comparison to FP measurements, a significant effect on the estimation of Vtoff ($p = 0.007$) and t ($p = 0.007$). The interaction between the MEN device and the point of attachment to the subject had, in comparison to FP measurements, a significant effect on the estimation of Vmax ($p = 0.009$) and Vtoff ($p = 0.003$). The interaction between the MEN system and the point of attachment to the subject had, in comparison to FP measurements, no significant effect on the estimation of t ($p = 0.72$). The interaction between the MYO system and the point of attachment on the subject had, in comparison to FP measurements, no significant effect on the estimation of Vmax ($p = 0.71$), Vtoff ($p = 0.64$) and t ($p = 0.55$).

The estimated Vmax had the best validity compared to the FP measurements, when the MYO system was attached to the subject's hip. The two-way ANOVA test showed no significant difference between the MYO and FP measurements for estimating Vmax ($p = 0.52$). No significant difference was observed between MEN and FP for estimating Vmax ($p = 0.89$) either. The best significant correlation for estimating Vmax ($r = 0.92$; $p = 2.39*10^{-8}$) was observed between the MYO system (attached to the subject's hip) and the FP (Table 1 and 2). The Bland and Altman test showed the best results when the MYO system was attached to the hip, demonstrating the best 95% limit of agreement for estimating Vmax in comparison to the FP (Figure 3). Low bias (> 0.06 $m.s^{-1}$), low reliability (± 0.22 $m.s^{-1}$) and good mean accuracy (< 0.1 $m.s^{-1}$) between the MYO system – when attached to the hip and FP – was observed for estimating Vmax.

When attached to the subject's back, the MEN equipment showed the best validity in comparison to the FP measurements. The two-way ANOVA test illustrated no significant difference between MEN and FP for estimating t ($p = 0.32$). A significant difference was observed between MYO and FP for estimating t ($p = 0.006$). The best significant correlation for estimating t was achieved (Table 1 and 2) between the MEN system when attached to the subject's back and the FP ($r = 0.97$; $p = 5.40*10^{-12}$). When attached to the back, the Bland and Altman plots of the MEN sensor presented the best 95% limit of agreement with the FP for estimating t (Figure 4). Low bias (< 0.018 s), similar mean accuracy (< 0.019 s) and very good reliability (± 0.028 s) between the MEN system attached to the back and the FP was observed when estimating t.

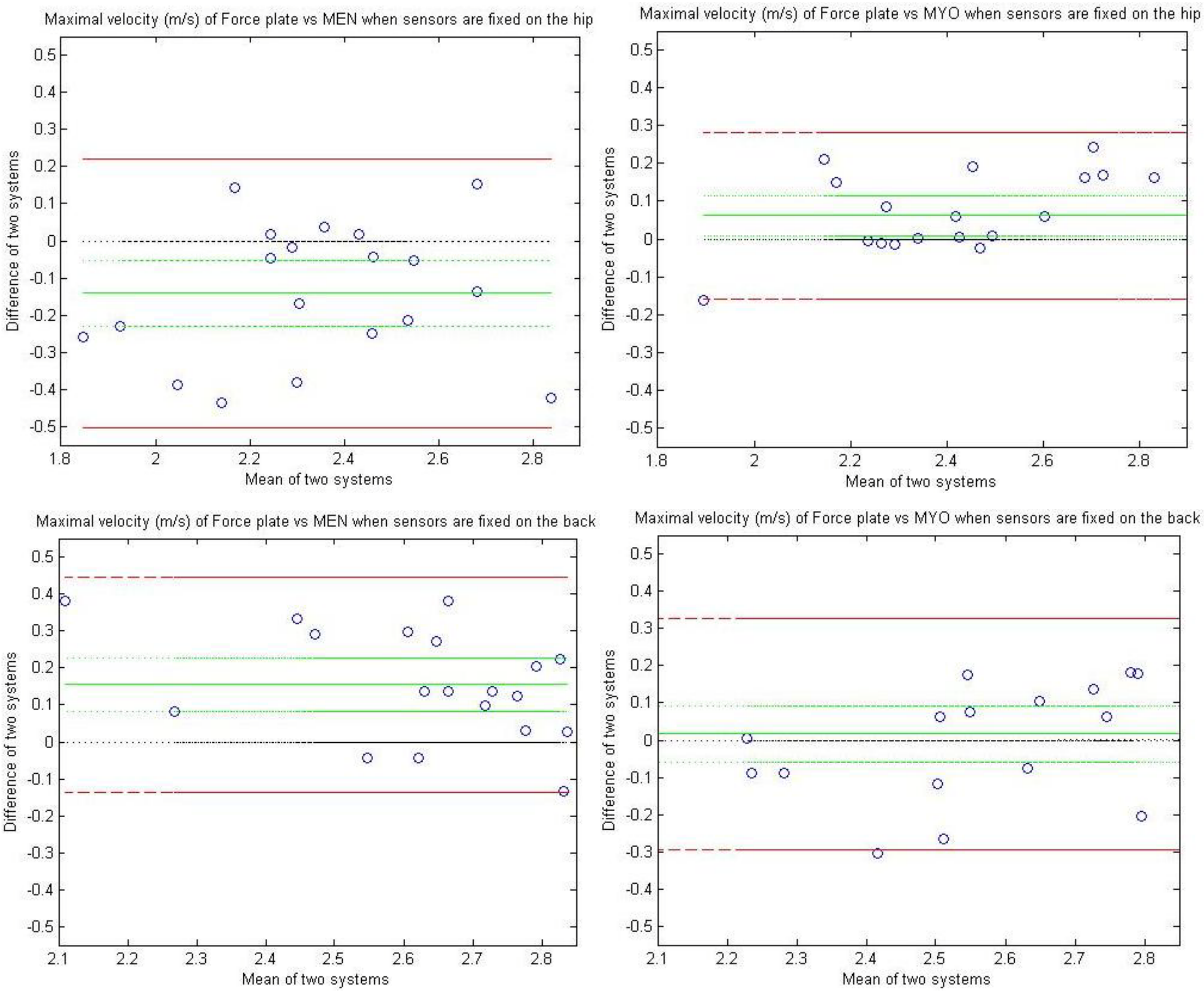

Figure 3. Bland and Altman plots of Vmax estimated by the MEN (left) and MYO (right) systems when attached to the hip (top) and back (bottom) vs. the FP measurements. The green lines represent the limits of bias whereas the red lines represent the 95% limits of agreement.

Table 1. Significant correlations between variables (Vmax, Vtoff, t) of the force plate and the MEN/MYO systems attached to the hip. * $p < 0.001$.

Variable	r value of MEN	p value of MEN	r value of MYO	p value of MYO
Vmax	0.776	$9.36*10^{-5}$ *	0.92	$2.39*10^{-8}$ *
Vtoff	0.758	0.0001 *	0.76	0.0001 *
t	0.858	$2.6*10^{-6}$ *	0.81	$2.55*10^{-5}$ *

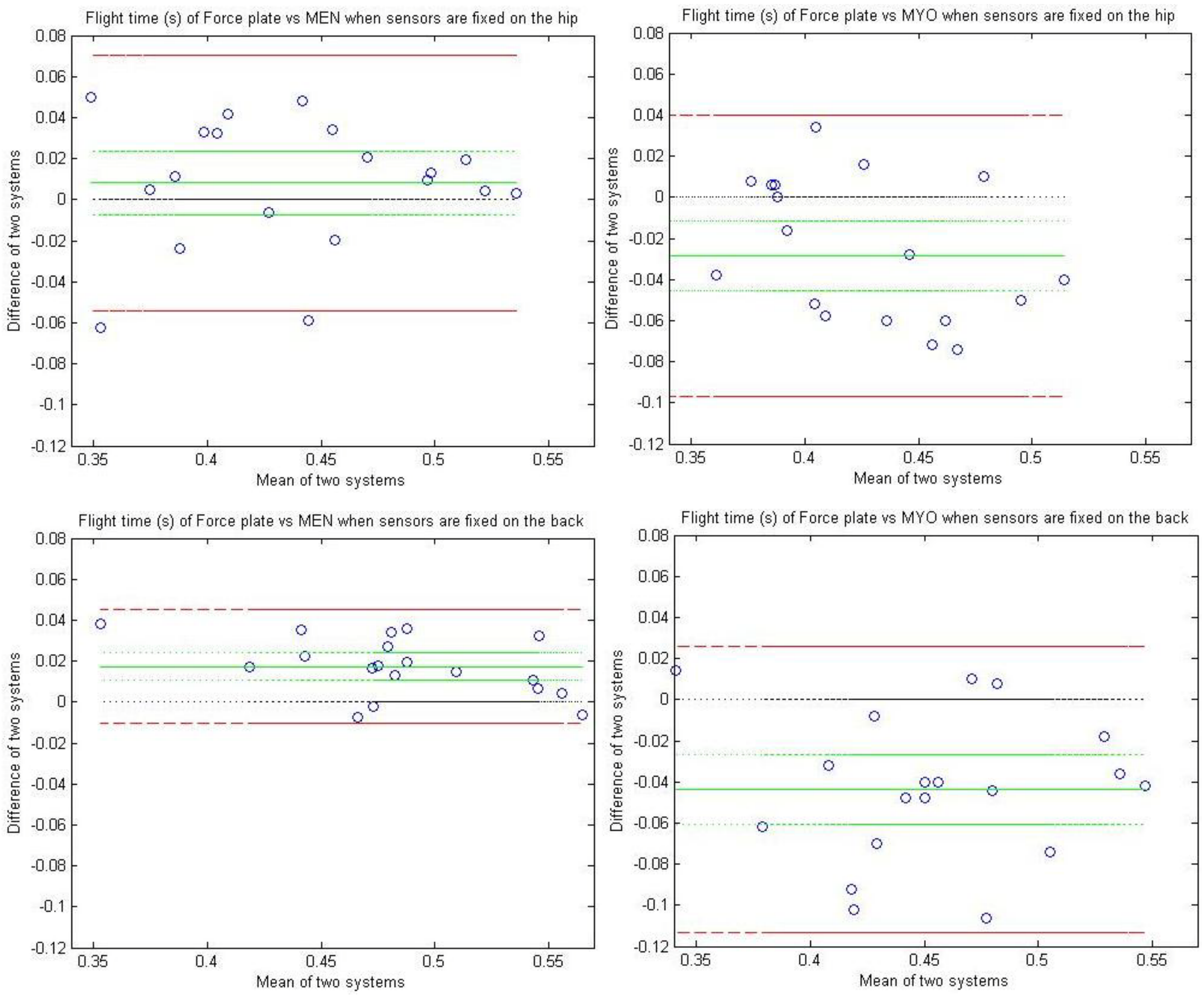

Figure 4. Bland and Altman plots of t estimated by the MEN (left) and MYO (right) systems when attached to the hip (top) and back (bottom) vs. the FP measurements. The green lines represent the limits of bias whereas the red lines represent the 95% limits of agreement.

Table 2. Significant correlations between variables (Vmax, Vtoff, t) of the force plate and the MEN and MYO systems attached to the back. * $p < 0.001$.

Variable	r value of MEN	p value of MEN	r value of MYO	p value of MYO
Vmax	0.774	$9.8*10^{-5}$ *	0.822	$1.57*10^{-5}$ *
Vtoff	0.761	0.0001 *	0.82	$1.68*10^{-5}$ *
t	0.97	$5.4*10^{-12}$ *	0.79	$4.24*10^{-5}$ *

The Vtoff values were estimated by both sensors with very low accuracy. The two-way ANOVA test showed no significant difference between the MEN and FP measurements for estimating Vtoff ($p = 0.27$). No significant difference was observed between MYO and FP for estimating Vtoff ($p = 0.32$) either. Significant correlations ($0.75 < r < 0.86$; $p < 0.001$) were observed for each point of configuration when estimating Vtoff using the MEN and MYO systems (Table 1 and 2). When attached to the back, the Bland and Altman plots of the MEN system presented the best 95% limit of agreement with the FP for estimating Vtoff (Figure 5). However, very low reliability ($\pm$ 0.30 $m.s^{-1}$), low mean accuracy (> 0.25 $m.s^{-1}$) and significant bias (> 0.15 $m.s^{-1}$) between the MEN sensor attached to the back and the FP could be observed.

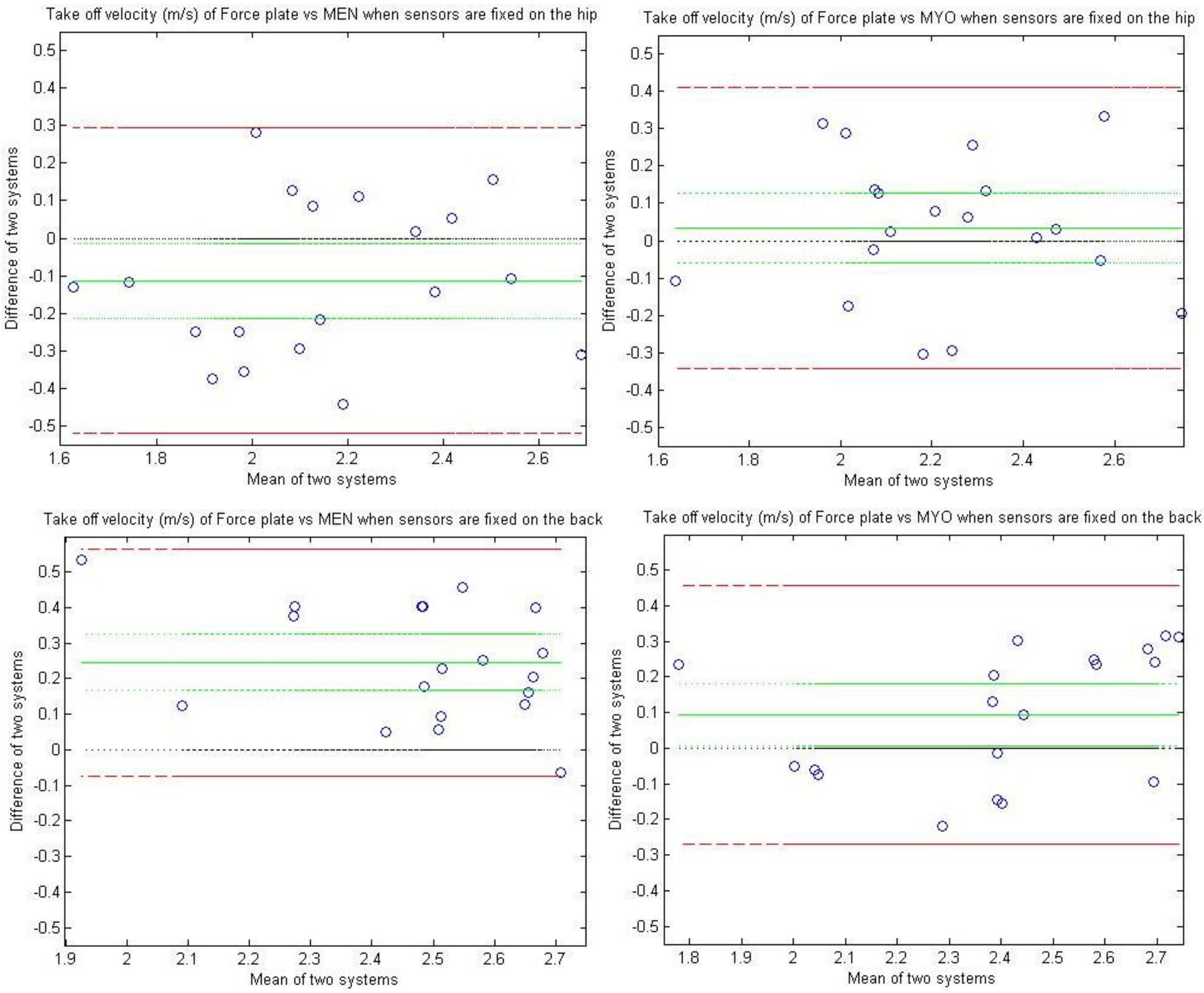

Figure 5. Bland and Altman plots of the Vtoff values estimated by the MEN (left) and MYO (right) systems when attached to the hip (top) and back (bottom) vs. the FP measurements. The green lines represent the limits of bias whereas the red lines represent the 95% limits of agreement.

Discussion

The present study showed that the point of attachment of the MEN and MYO systems influenced differently the estimation of the Vmax, Vtoff and t parameters. When attached to the subject's back, the MEN system estimated t with better accuracy than MYO in comparison to the FP. When attached to the subject's hip the MYO equipment estimated Vmax with better accuracy than MEN in comparison to the FP. Furthermore, the gathered Vtoff values from both, the MEN and MYO systems, are neither accurate nor reliable.

A limitation of this study was that the MEN system was fixed to the MYO equipment and not directly to the subject. This choice was made due to the assumption that when the MEN system is fixed to the MYO equipment, an exacter measurement of similar human motions can be achieved (since there is no movement between them). In this study, the MEN and MYO systems had similar noise risk assessment due to the same relative motion of the skin and the soft point of attachment. A second restriction of this research was the small number of jumps tested (n = 19 in each configuration). This can influence the outcome of the Bland and Altman plots and the statistical power of the study findings. However, the homogeneous population with homoscedasticity and normal distribution and the relative comparison of statistical results between them was investigated with the aim to compare the relative sensor accuracy and the influence of the point of attachment.

The obtained results showed that when attached to the hip, the MYO system was the most accurate and reliable for estimating Vmax of the subject's centre of mass during a squat jump in comparison to the FP measurements. This result is in agreement with the recommendations by Myotest SA. The computational methods could explain the differences between the results of both systems (MEN and MYO). While the procedure by MEN estimates the angular position of the sensor before and during the subject's motion, the computational method developed by Myotest SA seem to be more efficient to assess Vmax.

The present study showed very low accuracy and reliability for the estimation of Vtoff when using the MYO system in comparison to the FP. This result is in agreement with those of Casartelli et al. (2010), illustrating that the MYO system estimates Vtoff neither accurately nor reliably in comparison to photoelectric cells (Optojump). The authors state that the MYO system estimates the Vtoff parameter neither accurate nor reliable when applying numerical integration on the impulse phase during the subject's squat jump (Casartelli and al., 2010). In the present study, the MEN equipment showed better results in assessing Vtoff than the MYO sensors in comparison to the FP measurements. However, the MEN system was not accurate or reliable enough to assess Vtoff compared with the FP. The Vtoff results obtained from both sensors were influenced by the computational method. Another explanation might be also connected to the measurement techniques of the systems. There is evidence that if the centre of mass is used to evaluate squat jump performances (Kibele, 1998; Nagano et al., 2007; Bobbert et al., 2008; Samozino et al., 2008), accelerometers can only estimate the acceleration of the point to which it is attached. One possibility to reduce such problems is to choose a point that could better represent the centre of mass during squat jumps.

In addition, large correlations between the MYO system – when attached to the back – and the FP could be observed when estimating t ($r = 0.82$, $p = 1.68*10^{-5}$). This correlation is in agreement with the study of Casartelli et al. (2010). However, a significant difference ($p = 0.006$) and systematic bias were observed between MYO and FP for estimating t. The results by Castagna et al. (2012) show similar correlations ($r = 0.89$) and bias (-0.036 s ± 0.021). In the present study, the MEN system showed more accurate and reliable results than the MYO system, when attached to the back. Future investigations on the MEN equipment could help to examine fundamental laws of dynamics during the aerial phase as well as the flight time, thereby possibly improving the accuracy and reliability for determining the subject's take off velocity.

Conclusion

The current study suggested the use of commercial sensors for the purpose of assessing the flight time as well as take off and vertical velocities of squat jump performances. In comparison to the FP measurements, the used MYO equipment estimated Vmax with the best accuracy and reliability when attached to the hip. On the other hand, the MEN sensor assessed t with the best accuracy and reliability compared to the FP outcome when attached to the back. Both systems were inaccurate in estimating Vtoff and are therefore not interchangeable to the FP. As illustrated in the present research, the sensor's point of attachment to the subject as well as the implemented computational software had a direct influence on the kinetic variables of the squat jump performances. Since, in general, it is not possible to place the accelerometers exactly at the subject's centre of mass, further computational developments and detailed investigations regarding the subject's point of attachment have to be undertaken.

References

Atkinson, G. & Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sport medicine. *Sports Medicine, 26* (4), 217–38.

Bland, J.M. & Altman, D. (1995). Comparing methods measurement: why plotting difference against standard method is misleading. *The Lancet, 346*, 1085–87.

Bobbert, M.F., Casius L.J.R., Sijpkens I.W.T. & Jaspers R.T. (2008). Human adjust control to initial squat depth in vertical squat jumping. *Journal of Applied Physiology, 105* (5), 1428–1440.

Casartelli, N., Muller, R. & Maffiuletti, N. (2010). Validity and reliability of Myotest accelerometric system for the assessment of vertical jump height. *Journal of Strength and Conditioning Research, 24* (11)*,* 3186–3193.

Castagna, C., Ganzetti, M., Ditroilo, M., Giovannelli, M., Rochetti, A. & Manzi, V. (2012). Concurrent validity of vertical jump performance assessment systems. *Journal of Strength and Conditioning Research, 27* (3), 761–768.

Crewther, B.T., Kilduff, L.P., Cunningham, D.J., Cook, C., Owen, N. & Yang, G.Z. (2011). Validating two systems for estimating gorce and power. *International Journal of Sports Medicine, 32* (4), 254–258.

Dionyssiotis, Y., Galamos, A., Michas, G., Trovas, G. & Lyritis, G.P. (2009). Assesment of musculoskeletal system in women with jumping machnography. *International Journal of Women's Health, 1*, 113–18.

Dowling, J.J & Vamos, L. (1993). Identification of kinetic and temporal factors related to vertical jump performance. *Journal of Applied Biomechanics*, *9* (2), 95–110.

Garcia-Lopez, J., Peleteiro, J., Rodriguez-Marroyo, J.A., Morante, J.C., Herrero, J.A. & Villa, J.G. (2005). The validation of a new method that measures contact and flight times during vertical jump. *International Journal of Sports Medicine*, *26* (4), 294–302.

Glatthorn, J.F., Gouge, S., Nussbaumer, S., Stauffacher, S., Impellizzeri, F.M. & Maffiuletti, N.A. (2011). Validity and reliability of optojump photoelectric cells for estimating vertical jump height. *Journal of Strength and Conditioning Research, 25* (2), 556–560.

Hatze H. (1998). Validity and reliability of methods for testing vertical jumping performance. *Journal of Applied Biomechanics*, *14* (2), 127–140.

Houel, N., Faury, A. & Seyfried, D. (2010). Accuracy and reliability of the Mensense system to evaluate a squat jump. *Procedia Engineering, 12*, 3473.

Houel, N., Dinu, D., Faury, A. & Seyfried, D. (2011). Accuracy and reliability of the Myotest Pro system to evaluate a squat jump. *Procedia Engineering, 13*, 434–438.

Innocenti, B., Facchielli, D., Torti, S. & Verza, A. (2006). Analysis of biomechanical quantities during a squat jump evaluation of a performance index. *Journal of Strength and Conditioning Research, 20* (3)*,* 709–715.

Jidovtseff, B., Crielaard, J.M., Cauchy, S. & Croisier, J.L. (2008). Validité et reproductibilité d'un dynamomètre inertiel basé sur l'accélérométrie.[Validity and reliability of an inertial dynamometer using acceleromety]. *Science & Sports, 23*, 94–97.

Kibele, A. (1998). Possibilities and limitations in the biomechanical analysis of countermovement jumps: a methodological study. *Journal of Applied Biomechanics*, *14* (1), 105–117.

Lilliefors, H.W. (1967). On the Komogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62* (318), 399–402.

Nagano, A., Komura, T. & Fukashiro, S. (2007). Optimal coordination of maximal effort horizontal and vertical jump motions – a computer simulation study. *Biomedical Engineering Online, 6* (20), 1–9.

Nuzzo, J.L., Anning, J.H. & Scharfenberg, J.M. (2011). The reliability of three devices used for measuring vertical jump height. *Journal of Strength and Conditioning Research, 25* (9), 2580–2590.

Picerno, P., Camomilla V. & Capranica L. (2011). Countermovement jump percormance assessment using a wearable 3D inertial measurement unit. *Journal of Sport Sciences*, *29* (2), 139–146.

Quagliarella, L., Sasanelli, N., Belgiovinne, G. & Cutrone, N. (2006). Flying time evaluation in standing vertical jump by measurement of ankle acceleration. *Gait & Posture, 24*, 56–57.

Samozino, P., Morin, J.B., Hintzy, F., Belli, A. (2008). A simple method for measuring force, velocity and power output during squat jump. *Journal of Biomechanics, 41* (14), 2940–2945.

Optimizing Youth Baseball Batting Orders

Stephen J. Robinson

Department of Chemistry and Physics, Belmont University, Nashville, TN, USA

Abstract

Batting order (i.e., lineup) optimization for professional baseball teams has been analyzed for decades using various models and assumptions. In general, while optimization is useful – even with minimal benefit – it yields only small fractions of a run per game in improvement, mainly due to the interchangeability of most professional baseball players. Youth baseball, on the other hand, is a prime candidate for lineup optimization as it addresses large talent disparities and creates substantial improvements. In addition, a typical youth lineup is comprised of the same batters throughout a season, while this is not even approximately true for most major league teams; thus, finding "a lineup" for a youth team is more meaningful and useful in that sense. Here, the optimal lineup is considered to be the one which produces the highest number of runs per game. A probabilistic algorithm finds this optimal lineup by simulating many games with each possible lineup, allowing for more detailed statistical analysis. In addition, we see how run limits affect game outcomes and how individual players contribute to a team's performance. A study of a major league lineup is used for comparison.

KEYWORDS: BASEBALL, YOUTH, LINEUP, OPTIMIZATION

Introduction

For academics and fans alike, baseball lends itself quite well to the use of statistics. This is attestable to the discrete, one-on-one interactions of offense and defense that make record-keeping easy and quantifiable. For most of baseball's history, batting average, home runs, and runs batted in were standard measures of a hitter's prowess, while wins, strikeouts, and earned run average sufficed for pitchers. Recently, however, major league teams and fans have relied on computing power to quickly quantify a player's ability with statistics such as wins above replacement and even defensive measures such as range factor.

In addition to new statistics, computers are used increasingly to solve intractable predictive problems in sports. In baseball, one area that clearly lends itself to computational power is in determining the optimal batting lineup for a team. Lindsey (1963) was one of the first to propose calculating the average run value of different types of hits and using that to determine a player's usefulness in a lineup. Later, Pankin (1991) formalized this process in the form of Markov chain matrices that rely on the probabilities of transitions from one state (e.g., one out, runner on third) to another. Several others (Bukiet, Harold, & Palacios, 1997; Osawa & Aida, 2005; Sokol, 2006; Kakui & Arai, 2010; Graham, 2012) have expanded upon Pankin's work, while some have taken different approaches entirely, including Monte Carlo strategies (Freeze, 1973), graph theory (Sugrue & Mehrotra, 2007), and evolutionary computing (Chen, 2006). These methods assume the best lineups are found by maximizing the average number of runs scored per game (RPG). Hirotsu (2011) pointed out, however, that accounting for the standard deviation of RPG can lead to lineups that outperform those that simply maximize the average, while still acknowledging that the achieved gain is usually not worth the extra computation required.

Each of these studies has had as its goal the production of the best lineup for a professional baseball team. However, the vast majority of baseball teams are comprised of youth: in the United States alone, more than 11 million people play in a baseball league each year, and worldwide, more than two million youths play in the Little League organization in more than 80 countries and many more play in other leagues.

For these youth teams, there are generally many levels of stat-keeping, from zero to intermediate:

- In some games, and especially for very young players, the actual score of the game may not be kept at all.
- If the score is kept, it is often done by a volunteer parent who may only keep track of the runs for each team.
- If the most basic stats are kept (i.e., hits, strikeouts, and runs; if allowed, also steals and walks), it is usually without regard to the nuances of accurate score-keeping. For example, in a game between five-year-olds, a large percentage of "hits" should actually be scored as errors on the defense, thereby lowering, not raising, a batter's batting average. In addition, these stats are often simply left as raw data, as further regular computation of averages requires tedious manual calculations or knowledge of spreadsheets.
- In secondary school, more advanced statistics are often kept, but are generally no more complex than keeping track of hitting into double plays, stolen bases, and getting hit by pitches.

In addition to universally simpler statistics, youth teams, especially with children under ten years old, are different from professional teams in many other ways:

- There is a much wider range (i.e., standard deviation of statistical measures) of abilities.
- There are often run limits placed on a half-inning to keep games more competitive; when the limit is reaching, the half-inning is immediately over. This ensures that each team will get several chances to bat during a game.
- Time and inning limits may keep games between only four and six innings long.
- There may be more than 9 children (10-12 are common) batting in a given lineup.
- Tagging up on fly balls to the outfield is rare because they aren't generally caught.
- Batters may have statistics that are unheard of at higher levels (e.g., .800 batting averages).
- The hitters with high batting averages and the power hitters are usually one and the same. This makes the conventional major-league lineup – high-on-base-average hitters first, power hitters next, and weak hitters last – meaningless.
- Pinch hitters are rarely used.

There are far too many variables and combinations to address every possible playing situation, so the main thrust of this work is to provide general guidelines for youth baseball coaches to optimize their teams' lineups where statistical information is limited and where the conventional wisdom applied to major-league teams does not apply. It should be noted that the author has many years of experience coaching and playing for youth baseball teams, and the following assumptions arise from the most typical situations seen. It should also be noted that youth baseball can generally be divided into two categories. The first involves a coach pitching to his own team (coach-pitch); this is typically seen for players eight-years-old and under. The second involves players from the opposing team pitching to batters (player-pitch). Only in the player-pitch leagues do statistical categories such as the walk, hit-by-pitch, and stolen base come into play. To begin, we consider coach-pitch teams for the simplicity of explanation, and later consider older ages and more advanced statistics.

Methods

The runner advancement model created by D'Esopo and Lefkowitz (1977) is often used for determining how hitters and runners move around the bases. However, it is not fully applicable to youth teams, namely in the number of bases that runners advance on a given hit. Thus, the following assumptions apply to the calculations herein:

- Only the following statistics are kept: plate appearances (PA), singles (1B), doubles (2B), triples (3B), home runs (HR), and strikeouts (K). The base on which a player stands when the next batter comes up to bat determines the type of hit acquired; this eliminates the need to determine *how* the player got to his base, a necessity for youth stat-keeping.
- Runners already on base may move the number of bases given by a hit with a 50% chance of moving one extra base; runners do not tag up.
- The lead runner already on base has a 20% chance of being forced or tagged out.
- If a batter does not get a hit or a strikeout, he is out at first base; all other runners move up one base.
- A player will bat the same regardless of his position in the lineup and who bats before and after him.

Table 1. Statistics for a team of 4-6-year-olds over the course of 20 games.

player	PA	1B	2B	3B	HR	K
GB	64	32	8	7	6	6
CC	50	16	2	0	0	19
BG	55	17	4	0	0	28
LG	68	40	9	3	2	4
AH	61	25	9	0	0	15
JM	63	32	7	1	1	8
XM	61	36	6	1	0	7
HP	60	32	8	1	2	8
JR	67	33	10	4	8	2
XW	67	43	10	0	0	4
NW	66	33	13	3	5	5

The method used to find the optimal lineup here is similar to the ones used by Chen (2006) and Hirotsu (2011) in that each batter, based on his statistics, is assigned a probability of a given result (e.g., to get a double). Then, a random number generator ("the pitch") determines the result. Consider Table 1, which contains the statistics of the author's latest team of 11 4-6-year-olds through a 20-game season: the "Rays".

Consider JR, who has a probability of getting a single of 1B/PA = .493, and so on, as seen in Table 2. When JR comes "up to bat", the random number generator outputs a number $0 \leq x < 1$, giving a certain result as dictated by Table 2. If JR gets a hit, he becomes a runner, and his position is kept track of throughout an inning. The next batter then comes "up to bat", and the process is repeated throughout a given lineup. Runs and outs are tabulated for each inning until three outs are accumulated (or the half-inning run limit is reached), and the total number of runs for each game is kept track of as well. One particular lineup then "plays" many games (at least 100) to determine an average number of RPG. Since the number of innings per game is variable (see above), a complete simulation allows for a user-determined distribution of innings per game. In this article, we assume there are the same number of 4-, 5-, and 6-inning games played for each lineup.

Table 2. Using a random number generator, this table decides the outcome for player JR.

result	probability	random number x
1B	.493	$0 \leq x < .493$
2B	.149	$.493 \leq x < .642$
3B	.060	$.642 \leq x < .702$
HR	.119	$.702 \leq x < .821$
K	.030	$.821 \leq x < .851$
out at first	.149	$.851 \leq x < 1$

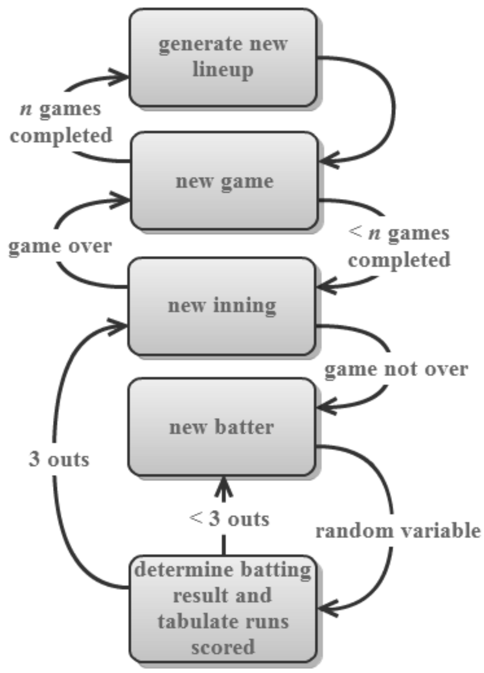

Figure 1. The method to determine the best lineup with n games per lineup. "3 outs" may also be reached with an inning run limit if applicable.

When a given lineup has completed its simulated number of games, we try the next possible lineup (using recursion in Java) and repeat. Note that the number of possible lineups for n players is $n!$: 9! ≈ 400 thousand, 10! ≈ 4 million, 11! ≈ 40 million, and 12! ≈ 500 million. Still, the computations are simple, and for a team of 11 players, a typical 400-game season taken over all possible lineups requires about 15 hours on a modern personal computer (which, if the algorithm were to be used by a coach with updated stats between games to find the best lineup, gives plenty of leeway). The optimal (or at least near-optimal) lineup is the one which produces the most RPG over the simulation period. This process is highlighted in Figure 1.

Results

One of the advantages of simulating games in this way instead of using Markov chains is the ability to calculate a standard deviation and median number of RPG to compare with real results. These are shown in for the Rays over 1000 games. With a 7-run limit per half-inning, the average lineup produced 12.4 RPG, while taking away this restriction brought the average up to 13.1 RPG. The best and worst lineups, as shown, performed substantially better and worse than this. In reality, over 20 games and with a 7-run half-inning limit, this team averaged 14.3 RPG with a median and standard deviation of 16 and 6.5 RPG respectively; their 2012 win-loss record was 15-5.

Table 3. The best and worst lineups for the 4-6 Rays for games in which a run limit exists or not. OPS is the addition of on-base and slugging percentages, a common measure of a hitter's performance.

	7-run limit per half-inning				**no run limit**			
	best lineup		**worst lineup**		**best lineup**		**worst lineup**	
	player	**OPS**	**player**	**OPS**	**player**	**OPS**	**player**	**OPS**
	GB	2.281	BG	.836	GB	2.281	BG	.836
	LG	1.897	JM	1.492	LG	1.897	XW	1.731
	JR	2.269	AH	1.262	JM	1.492	CC	.760
	XW	1.731	XM	1.541	XM	1.541	AH	1.262
	HP	1.700	CC	.760	HP	1.700	JM	1.492
	XM	1.541	JR	2.269	JR	2.269	GB	2.281
	NW	2.152	LG	1.897	NW	2.152	XM	1.541
	JM	1.492	HP	1.700	AH	1.262	JR	2.269
	BG	.836	NW	2.152	BG	.836	HP	1.700
	CC	.760	GB	2.281	XW	1.731	NW	2.152
	AH	1.262	XW	1.731	CC	.760	LG	1.897
mean	13.8		11.1		14.6		11.6	
median	14		11		14		11	
std dev	5.5		5.1		6.5		5.5	

Table 4 shows similar results for a 7-8-year-old team: the “Braves.”. With a 7-run limit per half-inning, the average lineup produced 11.2 RPG, while taking away this restriction brought the average up to 11.6 RPG. In reality, over 18 games and with a 7-run half-inning limit, this team averaged 11.2 RPG with a median and standard deviation of 12 and 5.1 RPG respectively; their 2012 win-loss record was 11-7.

Table 4. The best and worst lineups for the 7-8 Braves for games in which a run limit exists or not.

	7-run limit per half-inning				**no run limit**			
	best lineup		**worst lineup**		**best lineup**		**worst lineup**	
	player	**OPS**	**player**	**OPS**	**player**	**OPS**	**player**	**OPS**
	KG	3.019	JB	.700	MR	1.857	TE	1.409
	MR	1.857	EH	.946	JW	1.648	KN	1.373
	JW	1.648	TE	1.409	KG	3.019	JB	.700
	HS	2.130	GD	1.438	TE	1.409	JW	1.648
	AT	1.627	KN	1.373	BP	1.962	GD	1.438
	KN	1.373	JW	1.648	HS	2.130	EH	.946
	GD	1.438	KG	3.019	AT	1.627	BP	1.962
	BP	1.962	BP	1.962	EH	.946	AT	1.627
	JB	.700	AT	1.627	KN	1.373	MR	1.857
	EH	.946	HS	2.130	GD	1.438	HS	2.130
	TE	1.409	MR	1.857	JB	.700	KG	3.019
mean	12.5		9.9		13.0		10.3	
median	12		10		12		10	
std dev	5.2		4.6		5.9		5.1	

The large disparity in RPG between the best and worst lineups is something not seen in previous work (referenced above) with major-league teams. This is because the difference in abilities at the major league level is relatively minor, and switching players around produces minimal gain. At the youth level, however, the differences between players are much greater, and batting order becomes much more important.

In professional baseball, there is an unwritten rule of placing the worst batter at the bottom of the lineup so that he bats the least possible number of times in the game. At the youth level, however, because of the wide variability of player talent, it is beneficial, as shown in Table 3 and

Table 4, to place a decent hitter in the bottom spot so that he may be on base when the top of the lineup comes up again. Simulations show that the worst hitter often ends up a few spots higher – behind a power hitter – instead. It is our belief that this is because, with a half-inning run limit, the best hitters often bring runners home near the limit. Then the worst batter can get the last out of the inning, and the game can resume with a better hitter the next inning with the potential for more runs, the run limit having started over at zero. This theory is supported by the change of both the Rays' and Braves' lineups to place the worst hitter at the bottom when the run limit is removed.

In addition to the placement of the worst hitter, we see that conventional wisdom does not apply regarding power hitters either. Namely, at the higher levels of baseball, the 3-4-5 hitters generally have the highest slugging percentages. However, notice the placement of the three highest slugging percentages on the Rays: GB hits first, JR third, and NW seventh. On the Braves: KG hits first, HS fourth, and BP eighth. (GB and KG also have the highest on-base percentages on their teams.) From these results, it seems most effective to spread power out through the lineup, not concentrate it in a few spots together. Again, this makes sense given the stats of youth players. At the major league level, where teams average only about half a run an inning, it is always *unlikely* for a player to reach base, so runs should be scored when possible by putting the best hitters together. For youth teams, however, where run production per inning is around 2-3, it is usually more likely than not that a player will reach base, and a significant advantage is gained by loading the bases with players before letting the power hitters drive them in.

By examining the rate of occurrence of runs scored in each game the model simulates, we can see a predicted normal distribution as shown for the Rays in Figure 2. Given the relatively small number of games actually played, it is difficult to compare this to the real Rays, but a better comparison is shown later for a major league team.

One may ask whether youth teams have sufficient batting data over a short season to allow the model to be effective and/or accurate. Our experience in using this model for real teams has shown the answer to be yes. Mainly, youth players tend to show their above- or below-average abilities very early in the season and remain more consistent than professional players. This is most likely due to (1) the same pitcher pitching to all batters through the season and (2) work ethic generally not being a big part of a five-year-old's life; very young players tend to rely on natural ability much more than practice. Thus, relatively consistent lineups – not necessarily by name, but by general ability – are seen within just 3-4 games. After about a dozen games, the lineup will generally do no more than switch a couple of players with similar abilities.

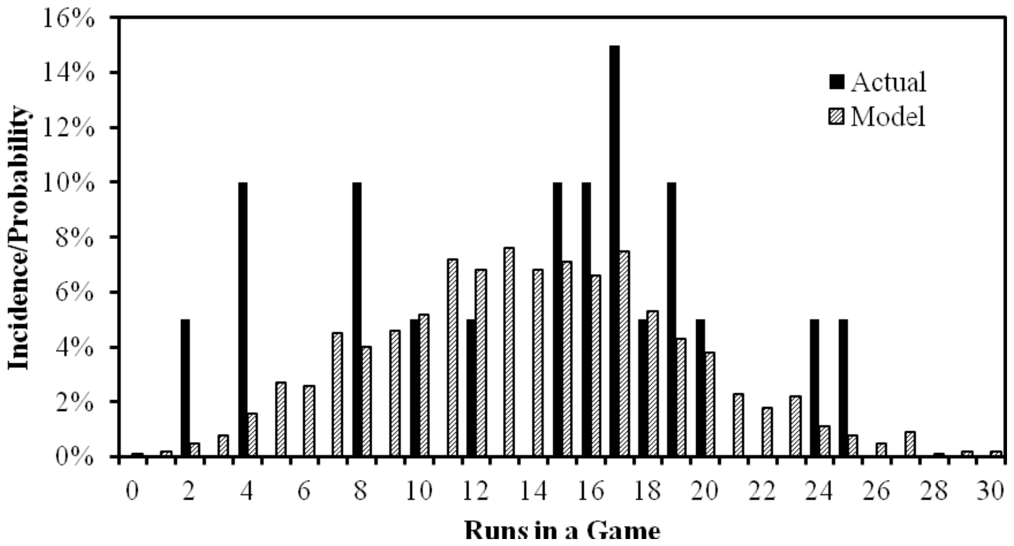

Figure 2. The distribution of runs scored in a game for the 4–6 Rays considering both the actual results and the model's predictions.

It is also interesting to see how the run limit per half-inning affects a team's RPG. As shown above, removing the seven-run limit only increases the RPG by between 0.5 and 0.8, so the average game length (or, more appropriately for youth, the number of innings played in a set time period) should not change drastically in doing so. Figure 3 shows that the run limit has the most impact on the game score – at least for the Rays and Braves – when it is less than seven. Interestingly, considering the improbability of a league commissioner performing the same simulation as seen in Figure 3 (or, for that matter, analyzing scorebooks to see the likelihood of a certain number of runs being scored in an inning), it is coincidental (and fortunate) that seven was chosen as the half-inning limit; it has a relatively small effect on the total runs scored while protecting less-capable teams.

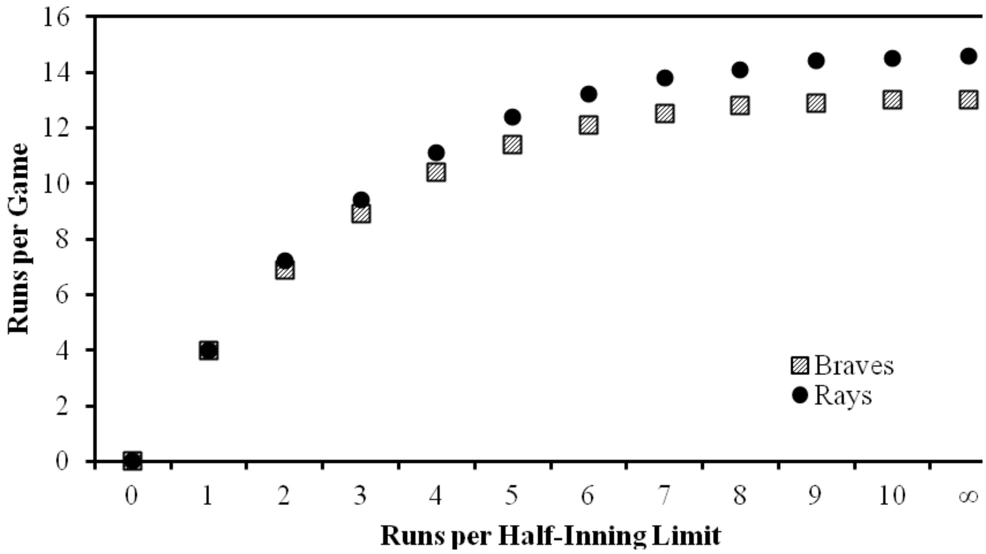

Figure 3. RPG for the Braves and Rays considering the maximum number of runs allowed per inning.

In terms of winning a baseball game, defensive statistics are clearly important, but holding them constant, we can easily find the increase in expected wins per season using this model by calculating the chances of beating an average opponent assuming a normal distribution of runs for both teams. Similar work was performed by Stern (1991) to predict a football team's seasonal wins. Specifically, the chances of winning are found by finding the cumulative probability of a normal distribution for $x > 0$ with a mean equal to the difference of the two teams' RPG means and a variance equal to the sum of the teams' RPG variance. The Rays' opponents averaged 9.6 RPG with a standard deviation of 5.9. For the Braves, it was 8.1 and 6.3. Using this method, the worst and best Rays' lineups will beat their opponents 58% and 70% of the time. For a 20-game season, this is a difference of 2.4 wins. For the Braves, probabilities range from 59% to 70% for 2.2 wins. For a 162-game season, this would equate to about 19 wins. Previous work has shown that professional baseball lineup optimization may only increase win expectations by about one game per 162-game season. (Tango, Lichtman, & Dolphin, 2007).

Individual Contributions

A coach may be interested in how an individual player contributes to his team's overall success. As mentioned above, OPS is widely considered to be a good measure of a hitter's ability. To see how youth OPS is correlated with run production, each player was removed from the lineup and a new optimal lineup and expected RPG were calculated. Then, the player's relative OPS was compared to the team's improvement in run production when he is in the lineup (assuming only optimal lineups). Figure 4 shows the results. As an example, consider BP. His OPS is 17.7% higher than the team average excluding BP. When he is in the lineup, the team's RPG increases by 3.3%. (Note that this is only possible for youth teams, as they typically play with more than nine players and one can be removed from a lineup without being replaced.) In terms of RPG, this can be significant – for the Rays, the changes in run production vary from a decrease of 1.5 RPG to an increase of 1.1 RPG for an individual player, and between -1.4 to 1.8 RPG for the Braves. Again, at the major league level, the players are more evenly talented and the RPG is so few that the contributions of a single player are far less noticeable. But for a youth coach who might have to make a decision to sit a player because of defensive restrictions, this information is quite valuable. On average, a player's OPS percentage difference (positive or negative) yields about one-fifth of that in run-production percentage change, as shown in the linear regression. The offset of 0.01 indicates an average RPG increase of about 1% if any player is removed the lineup and is indicative of the fact that 55% of the players on the Rays and Braves had an OPS below their team's average.

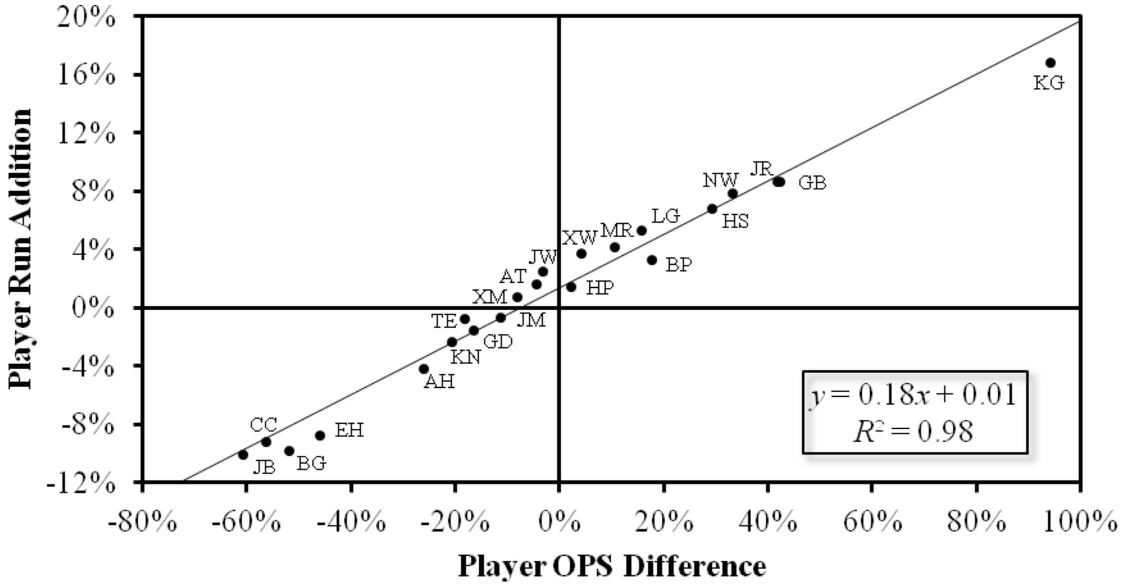

Figure 4. The correlation between each 4-6 Rays and 7-8 Braves hitter and his contribution to his team's RPG. "Player OPS Difference" is the percentage that a player's OPS is above or below the rest of the team's combined OPS. "Player Run Addition" is the percentage increase or decrease of a team's runs when the player is in the optimized lineup vs. when he is not.

The effect of individual player performance can also be translated into increasing the effective number of wins for his team. In professional baseball, wins above replacement (WAR) is most often used for this purpose (Goldman, 2012), but it includes defensive and other statistics that are not available at the youth level. Using the same method described above to calculate a team's expected increase in wins, we can find how many wins a player is expected to earn for his team based on his OPS difference alone. The result is that, on average, 0.104 wins (for a 20-game season) are generated by each player for each percentage point of run addition added in Figure 4. For example, the 7-8 Braves' KG would earn 1.60 wins, which, when translated to a 162-game season, becomes 13. WAR compares players with other players, while this statistic compares a single team with or without a given player in the lineup. The highest ever offensive WAR (oWAR) in the major leagues belonged to Barry Bonds in 2001 at 12.2. Given the wild disparity between youth and major league statistics, the number of expected wins generated seems one of the most reasonable ways to compare individuals at different levels.

Advanced Statistics

As mentioned earlier, younger teams either do not or cannot keep track of more advanced statistics. If the method discussed thus far were to be used at higher levels of play, we could take advantage of other probabilities: the number of times stealing a base (SB), getting caught stealing (CS), and grounding into a double play (GDP) take a player's speed into account, while walking (BB) and getting hit by a pitch (HBP) quantify a player's ability to simply get on base.

We must use these more advanced statistics in conjunction with different assumptions than we made for youth teams. The following are different:

- The lead runner already on base has a 5% chance of being forced or tagged out if the batter gets a hit.

- Runners have only a 50% chance of reaching the next base when the batter is “out at first,” which includes both having a fly ball caught and being forced out at first base. Sacrifice hits are not accounted for because they are highly dependent on other hitters being on base, which would nullify the assumption that batters are independent of each other.
- During another player’s at-bat, a runner will attempt a steal of an open second or third base with probability (SB + CS)/ (1B + BB + HBP + SB + 2B). The denominator is an estimate of the number of times a runner stands on first or second base in a season. Adding SB overestimates the number of times on second (since there are steals of third as well), but the model does not take into account other methods of reaching first base (e.g. errors, fielder’s choices, etc.), which provides a complementary underestimation. Since runners are on first base much more than second base, there will be far fewer steals of third base, which mimics real baseball. Note also that (SB + CS) automatically accounts for the real-life probabilities of the next base being occupied, thereby forbidding a steal.

To test our model’s ability to predict an optimal lineup at the major league level, we tried it on the 2012 Chicago White Sox; their statistics are shown in Table 5.

Table 5. Batting statistics for the 2012 Chicago White Sox.

name	PA	1B	2B	3B	HR	K	BB	HBP	GDP	SB	CS
Beckham	582	83	24	0	16	89	40	7	10	5	4
De Aza	585	103	29	6	9	109	47	9	1	26	12
Dunn	649	50	19	0	41	222	105	1	8	2	1
Konerko	598	111	22	0	26	83	56	7	16	0	0
Pierzynski	520	84	18	4	27	78	28	8	8	0	0
Ramirez	621	120	24	4	9	77	16	4	15	20	7
Rios	640	114	37	8	25	92	26	4	18	23	6
Viciedo	543	85	18	1	25	120	28	6	18	0	2
Youkilis	509	67	15	2	19	108	51	17	10	0	0

The 5000-game model predicted the optimal lineup in

Table 6. Only nine-inning games were simulated, because even though a substantial portion of a team’s games allow them only eight innings to bat (i.e., they are winning in the ninth inning at home), there are also a number of games that go extra innings (potentially forever); a reasonable compromise is nine innings.

Table 6. The most frequent actual lineup for the Chicago White Sox in 2012, along with the best and worst predicted lineups.

position	real order	OPS	best	worst
CF	De Aza	.760	Youkilis	Ramirez
3B	Youkilis	.745	Viciedo	Rios
DH	Dunn	.797	Pierzynski	Viciedo
1B	Konerko	.857	Beckham	Youkilis
RF	Rios	.850	Konerko	Beckham
C	Pierzynski	.827	Dunn	De Aza
LF	Viciedo	.744	Rios	Konerko
SS	Ramirez	.651	De Aza	Pierzynski
2B	Beckham	.668	Ramirez	Dunn
mean	4.6		4.6	4.1
median	4		4	4
std dev	3.2		3.0	2.8

Note that the best predicted lineup is quite different than the actual most frequently used lineup, but the seasonal RPG and run distribution (seen in Figure 5) are very similar – for 0-8 runs in a game, the model gives χ^2 = 5.65 and p = 0.77. (Including the outlier of 9 runs in a game increases χ^2 to 10.2, giving p = 0.42.) This is a prime example of the interchangeability of most major league players – in 2012, 74% of major league hitters had WAR (wins above replacement) values between -1.0 and 1.0 (fangraphs.com). Note also that the continuity of the 2012 White Sox lineup is rare, in that eight of these nine players were eligible (i.e., had enough plate appearances) for a batting title using only their White Sox stats; in the vast majority of cases, finding a consistent optimal lineup for professional players is futile.

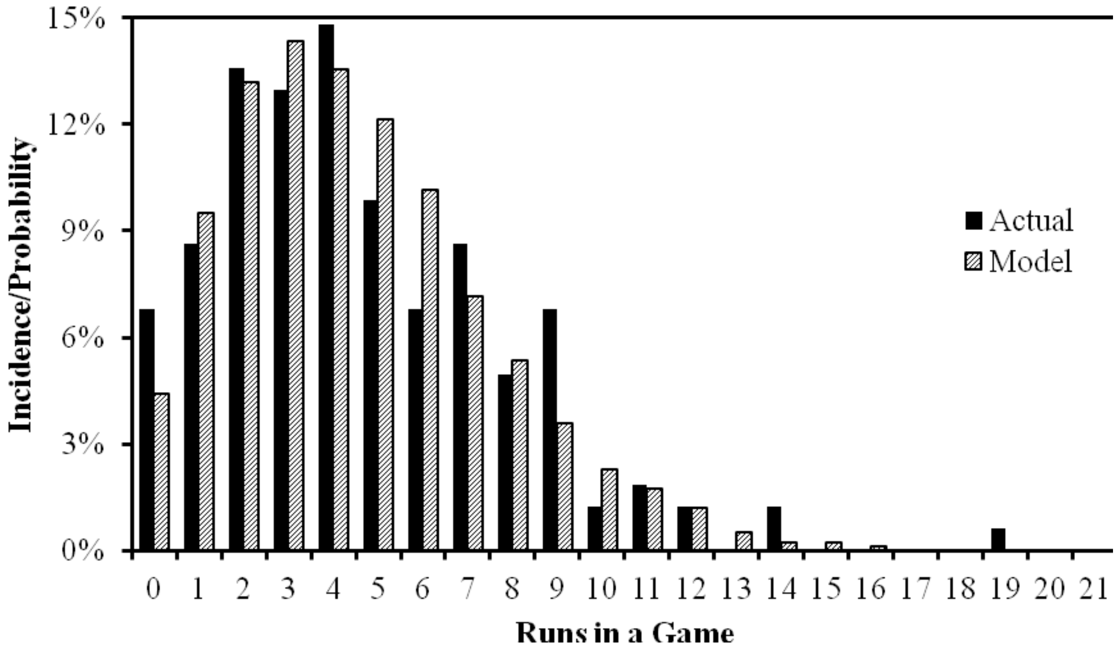

Figure 5. The distribution of runs scored in a game for the Chicago White Sox considering both the actual results and the model's predictions.

Conclusion

It is often meaningless to speak of "a lineup" for a professional baseball team since so many players are traded or injured during a regular season. That is, the players available in the beginning, middle, and end of a season would almost certainly be quite different, requiring game-by-game optimization. It remains a good strategy, but given the small benefit coupled with the major league egos that might resist change, one might question the work required to generate each lineup.

Youth baseball lineups – which far outnumber professional lineups – are much better suited for optimization for several reasons. (1) The net gain in RPG is far greater, making the advantage much clearer. (2) The same players play every game, allowing for consistency and removing the need to continually rotate players in and out of the lineup. (3) Youth coaches are generally less knowledgeable about baseball than professional managers so that computer-assisted optimization can create a substantial advantage over one's opponent. (4) The time between games allows for plenty of computation. (5) The necessary statistics for good results are simple and easy to understand.

This paper has detailed the use of probabilistic at-bats to determine basic guidelines for youth baseball lineups. The most intriguing (i.e., those which contradict conventional wisdom) are:

- If a half-inning run limit is in place, the last batter should be an average hitter, not the worst hitter. The benefit added by a good last-batter/first-batter interaction overcomes the fewer at-bats this batter receives.
- Power hitters should not be bunched together; but rather spread out over the lineup to drive runs in.

In addition, the following information was verified:

- The run contribution of a player is directly proportional to his OPS contribution; it does not seem to level off at either end of performance.
- The half-inning run limit seems optimized for game play at around seven runs.
- Fairly accurate modeling of professional teams can be done with a small set of simple statistics.

These general principles should prove useful to coaches and parents interested in youth baseball.

References

Bukiet, B., Harold, E.R. & Palacios, J.L. (1997). A Markov Chain Approach to Baseball. *Operations Research*, *45*, 14–23.

Chen, Y.-Y. (2006). Batting Order Optimization Using Evolutionary Computation. *Proceedings of the fourteenth International Conference on Genetic and Evolutionary Computation Conference Companion* (pp. 599–602), Montreal, Canada.

D'Esopo, D.A. & Lefkowitz, B. (1977). The Distribution of Runs in the Game of Baseball. *Optimal Strategies in Sports* (edited by S. P. Ladany and R. E. Machol, Eds., North-Holland Publishing Co., Amsterdam), 55–62.

Freeze, R.A. (1973). An Analysis of Baseball Batting Order by Monte Carlo Simulation. *Operations Research*, *22*, 728–735.

Graham, C.J. (2012). Baseball Enigma: The Optimal Batting Order. *MIT Sloan Sports Analytics Conference*, Boston, USA.

Goldman, S. (2012). *Extra Innings: More Baseball Between the Numbers from the Team at Baseball Prospectus*. New York: Basic Books.

Hirotsu, N. (2011). Reconsideration of the Best Batting Order in Baseball: Is the Order to Maximize the Expected Number of Runs Really the Best? *Journal of Quantitative Analysis in Sports*, *7* (2), 13.

Kakui, Y. & Arai, S. (2010). Scene Evaluation of a Ball Game for Solving Batting Order Optimization. *SICE Annual Conference 2010* (pp. 3250–3257), Taipei, Taiwan.

Lindsey, G.R. (1963). An Investigation of Strategies in Baseball. *Operations Research*, *11*, 477–501.

Osawa, K. & Aida, K. (2005). Speed-up Techniques for Computation of Markov Chain Model to Find an Optimal Batting Order. *Proceedings of the Eighth International Conference on High-Performance Computing in Asia-Pacific Region* (pp. 315–322), Beijing, China.

Pankin, M.D. (1991). Finding Better Batting Orders. New York: SABR XXI.

Sokol, J.S. (2003). A Robust Heuristic for Batting Order Optimization Under Uncertainty. *Journal of Heuristics*, *9*, 353–370.

Stern, H. (1991). On the Probability of Winning a Football Game. *The American Statistician*, *45* (3), 179–183.

Sugrue, P. K. & Mehrotra, A. (2007). An optimisation model to determine batting order in baseball. *International Journal of Operations Research*, *2* (1), 39–46.

Tango, T.M., Lichtman, M.G., & Dolphin, A.E. (2007). *The Book: Playing the Percentages in Baseball*. Washington, D.C.: Potomac Books.

Tactics Analysis in Soccer – An Advanced Approach

Jürgen Perl[2], Andreas Grunz[1] & Daniel Memmert[1]

[1]German Sport University, Cologne, Germany

[2]Johannes Gutenberg-University, Mainz, Germany

Abstract

In order to run a game tactically, high level knowledge is required from and by coaches and analysis experts. Assessing tactical performance through statistic indicators has some drawbacks, however, it is not only difficult to prove reliability on the defined indicators, but those static indicators often hide the game's dynamics. New network-based approaches offer a promising way to improve future evaluation of tactical performance and recognition of game dynamics. The aim of this article is to introduce a new approach where pattern-based tactics analysis is combined with success-oriented statistical frequency analysis. Therefore, the neural network-based pattern analysis of the SOCCER-approach (Perl & Memmert, 2011) has been completed by an event-oriented statistical analysis, which is mainly based on an automatic recognition of ball possession as an indicator of success. After a short introduction into basic aspects of game analysis, automatic position tracking and net-based pattern analysis, the new concept of SOCCER is presented in two steps: The first part deals with net-based analysis of dynamic processes, oriented in constellations of tactical groups of players. The second part deals with rule based semantics analysis, which allows automatic recognition and evaluation of individual activities and embedding them into the tactical patterns - thus enabling both, evaluation of tactical processes based on individual success as well as evaluation of individual activities in the context of tactical processes.

KEYWORDS: COMPUTER SCIENCE, GAME ANALYSIS, SOCCER, PATTERN RECOGNITION, NEURAL NETWORKS

Introduction

In the quarterfinal of the FIFA World Cup 2006, Germany versus Argentina, the goalkeeper of the German national squad, Jens Lehmann, paved the way into the semi-final. Information about shooting preferences of his opponents, gained by the analysis of their penalty striking habits, helped Lehmann in the decisive shootout to choose the right corner and to perform saves (Buschmann & Nopp, 2006). At the FIFA World Cup 2010 it was determined that the opponents of the English squad operated with an increasing number of long passes, which revealed deficits in the English defense (Buschmann & Nopp, 2010). In the round of sixteen in the match against England, the German squad took advantage of these findings. The goalkeeper Manuel Neuer made an assist by passing a goal-kick over the English defense line to forward Miroslav Klose who scored the lead. The examples highlight how performance analysis in soccer can be a key factor in effective game preparation, in addition to the monitoring and evaluation of training (for a review, Nopp, 2012).

On the one hand, spectacular cases like those are impressive. They simplify how important it is to get key information out of the game. On the other hand, they demonstrate the large gap between isolated key information and the knowledge on complex tactical behaviour. In order to recognize tactical plans and their success, it is necessary, based on recordable data, to recognize and analyse the behavioural patterns of the players and in particular of the tactical groups. Only if it is understood, in which interaction what constellation or move of a tactical group is successful and opens ways for individual tactical maneuvers, the tactical orientation of the team and the individual players can be improved.

Since the early 1970s, computerized notational analysis has been playing an important part as a scientific basis for developing concepts of recording and analyzing data from games. There are some scientific studies that improved the understanding of soccer by underpinning soccer performances analysis with theoretical findings (Memmert & Perl, 2009b, for a review). Hughes & Franks (2005), for example, clarified the need for normalization of data – the latter research group counted the number of passes per possession and linked that to resulting goals; due to unequal occurrences of each possession length (according to passes) a normalization of the data was inevitable. Therefore, Hughes & Franks (2005) divided "the number of goals scored in each team possession by the frequency of the sequence length" (p. 511), multiplied the results by 1000 in order to avoid small ratios – and by doing so, highlighted differences between successful and unsuccessful teams considering the style of play (possession play or direct play) and resulting conversion ratios to shots on the goal. This was based on another far-reaching study conducted in 1968 by Reep & Benjamin. This study coined the British and Norwegian soccer team's style of play by arguing that direct play – implying few passes and a high frequency of shots on the goal; thus playing highly penetrative – results in successful outcomes.

Obviously, such data is important to receive information about a lot of performance indicators reaching from technical skills over physical condition to most relevant skills like the ability of scoring goals.

However, most of those distributional and frequency oriented results, although doubtlessly helpful in practice, are primarily useful for classification and ranking, but neglect the dynamic aspects of processes like interaction and context. They "freeze" the ninety-minutes-game to just some handful of numbers, helping to understand "what" but not "why". The key for a better understanding of the game is to analyze what the coach is doing: He "reads" the game, i.e. he does not record numbers but recognizes patterns.

Therefore, the aim of this article is to give an introduction into methods and first results of pattern-based tactic analysis, which has been successfully run by means of artificial neural networks approximately during the last 4 years. Although it is not perfect at all, it shows that those networks can help to map the complex game to a sequence of patterns. Moreover, those patterns can be combined in a fruitful way with statistical results and/or the patterns themselves can be used for statistical analyses.

In the following, the emphasis is put on three main aspects:

The best fitting data for description and analysis of spatio-temporal processes like those in soccer is given by the positions of the players and the ball. Therefore, the first part deals with automatic position data recording. A brief overview introduces the most interesting up-to-date approaches.

In order to analyse the game processes in a qualitative way, the focus is laid on patterns of actions and interactions. Some main aspects of pattern analysis, in particular by means of artificial neural networks, are briefly introduced in the second part.

The third part introduces a net-based software-tool "SOCCER", which combines net-based pattern analysis with rule-based semantics, statistical analyses and expert knowledge to gain greater insights into the game dynamics. Moreover, it is very important to make data easily understand considering the data output. SOCCER, like most of the analysis software systems available, has integrated graphics, which help in terms of data interpretation. Whether the performance is coded in-event or post-event, the data is reinterpreted based on edited videotapes or graphical illustrations calculated by the software. Modern systems make it possible to refine actions in different areas of the pitch or at certain intervals of the game, e.g. during a certain match status. Finally, data can be used to compile statistics that can help to identify performance profiles of players and/or teams as well as the strengths and weaknesses.

Position Tracking in Soccer

In computerized notational analysis, distances covered during a game are of interest to quantify the intensity of motion. The amount of walked, jogged and sprinted meters as well as the velocity and acceleration of each player during a game are indicators for the condition of the team. In order to analyze actions and interactions, first of all the constellations of the players and the types of moves are of interest in computer-based tactics analysis. At any case, the positions of players and the ball play the important role and therefore have to be recorded from the game.

Given that video tracking is available in the majority of competitive games, research has focused on methods to extract position data of the players from recorded videos. In 1990, Herzog & Retz-Schmidt proposed a tracking system using image processing. They used a fixed camera that covers the entire pitch to record image data. This is a software programme that simply instructs the computer to take the current position, in (x, y)-pixels on the screen at any frame, and to calculate the distance travelled since the (x, y)-pixels at the previous frame. A first commercial prototype, the computer vision system AMISCO, was released in 1998 by Videosports Ltd. to obtain position data. Since then, the method of image processing made a big jump ahead, however, it still needs a supervisor to control and correct data. Impellizzeri, Sassi & Rampinini (2006) showed that the variation of different supervisors is below 2% in one tracked game.

More recent systems are based on frequency modulated continuous wave (FMCW) technology to track positions. The most common representative of this technology is the GPS system. Players have to carry a transmitting unit while playing, to be detected by the system. A comparison of accuracy of GPS and video based systems showed an overestimation of 4.8% for GPS and 5.8% for video examining distances during games (Edgecomb & Nortona, 2006). At present, the LPM system produced by ABATEC has the lowest distinction of just 1.6% underestimation of real and measured distance (Stelzer, 2004). It uses FMCW sensors that send a permanent signal to a certain number of receivers around the pitch. A central processing unit collects these data and is able to present it in real time. Although FMCW systems a more accurate and faster in presenting the data, they can just be used in training due to FIFA regulations at this time.

The position data collected with the different approaches in particular can be used to fulfill net-based tactics analysis, presented in the following.

Pattern Recognition

Pattern recognition based on position data can result in different kinds of patterns (Memmert & Perl, 2009a):

The formations of (tactical) groups of players – i.e. the positions of the group members in relation to each other – build spatial patterns. The time-depending movements of such groups build temporal patterns. Both types of patterns help to recognize tactical concepts. In combination, these patterns build spatio-temporal patterns of the game processes. And the combination of respective patterns of both teams result in interaction patterns, which are helpful in order to measure the success of tactical actions in the context of tactical interaction.

The patterns, if once obtained, can be used to calculate several statistics like frequency or rareness of action or interaction patterns. In combination with additional semantic information like success/failure, this can already give deeper insights into the game dynamics. In our research we could identify several rare patterns which also had a high probability of success. The terms "success" and "failure" can be defined in several ways depending on the research question (compare the section "Rule based semantics analysis").

To obtain such patterns, a lot of methods are available, stretching from simple similarity analysis over statistical clustering methods to neural network approaches. Basing on a lot of positive experiences with sports as well as with technical applications, we decided to use neural networks – in particular because in the case of self-organizing maps (see explanations below), no pre-information about number and types of clusters is necessary.

The applied methods can be divided in supervised and unsupervised methods. The term supervised/unsupervised has its origin in the scientific field of machine learning. Supervised methods learn patterns by examples. For instance the group tactic *wing play* in soccer can be learned by feeding the net with examples of *wing plays*. In contrast, unsupervised methods are characterized by the lack of labeled data. A similarity measure is used to group the data and to construct distinct prototypes. While in supervised learning a predetermined pattern is learned by examples, in unsupervised learning there are no predefined patterns. Therefore, the patterns gained by unsupervised methods do not necessarily correspond to conventional standard patterns like *wing play*, but surprisingly often do: If a pattern is sufficiently represented by the data and distinct from other ones, the net will normally recognize it by its own.

The self-organizing map (SOM) developed by Kohonen (2000) is an artificial neural network of the unsupervised type. A SOM consists of a set of artificial neurons that are connected to each other through edges usually arranged in a rectangular grid. During the training phase the network adapts itself to the distribution of the data used for training. After training each neuron encodes a different pattern. In an additional step, the neurons encoding similar patterns are grouped into clusters, representing a pattern-prototype.

The kind of patterns represented by the network is determined through the data that is used for training. If the network is trained with movements of one group of players, the resulting patterns will encode typical movements of that group. If the network is trained with interactions, consisting of movement data from 2 interacting groups and the ball, the resulting patterns will encode typical interactions.

The Dynamical Controlled Network (DyCoN; Perl, 2004), which is derived from SOM and overcomes several technical limitations of the original SOM-concept, has successfully been used to detect tactical patterns in soccer games (Memmert & Perl, 2009a, 2009b; Grunz, Memmert & Perl, 2012).

An example for supervised learning of specific tactical patterns in soccer like game initiations is given in Grunz et al. (2012). A hierarchy of neural networks was developed to learn these patterns. As the hierarchy is trained with labeled data gained by an expert watching and categorizing the game, the hierarchy belongs to the supervised methods. Several sets of example data extracted from the categorization were used to learn the corresponding patterns. For instance, the expert categorized a set of sequences that showed short game initiations. The position data corresponding to these sequences was extracted by a program to train the hierarchy of networks. After training the hierarchy, it was used to classify tactical patterns in new games. While an expert needs at least 5 hours to manually categorize a game, the developed approach can reduce the required time to a few minutes.

Combination of Pattern Recognition with Rule-Based Semantics and Statistical Analysis

This net-based game analysis approach in soccer is mainly based on the ideas of data reduction (Perl, 2008), pattern recognition (Grunz, Memmert & Perl, 2009; Memmert & Perl, 2009a, 2009b; Grunz, Endler, Memmert & Perl, 2011), the network DyCoN (Perl, 2004; McGarry & Perl, 2004), and the special game analysis software SOCCER (Perl & Memmert, 2011), which all work together as it is briefly described below.

The process starts with position data preparation and pre-processing, which is done by means of the software tool SOCCER, followed by three steps of analysis:

(1) The position data of the players of a team are reduced to those of tactical groups like offense or defense, followed by normalization to standard patterns, as it is shown in Figure 1.

(2) The net is trained with those formations, resulting in a collection of formation clusters, each containing a collection of variants of the corresponding formation type.

(3a) Along the time-axis, position data of interacting tactical groups are fed to the net, which recognizes the time-dependent corresponding formation types as well as in particular striking features.

(3b) Additionally quantitative analysis of frequency distributions of formation types is done by means of the statistics tool of SOCCER.

(3c) The trajectory analysis component of DyCoN enables tactical analyses of the game, including interaction and phase analyses. In particular long term interaction patterns as well as hidden or creative tactical activities can be recognized and analyzed regarding success.

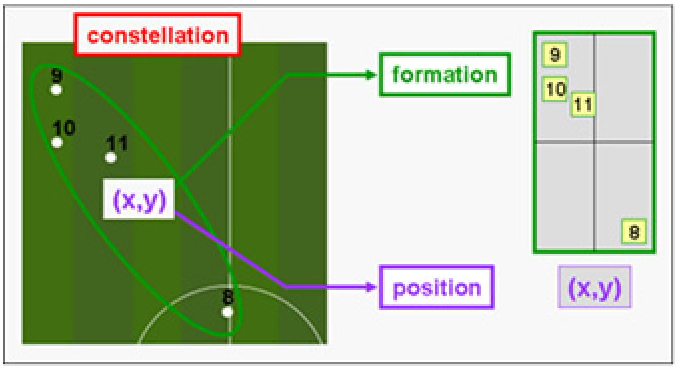

Figure 1. Departing a constellation into its position (centroid) and its characteristic formation (taken from Perl & Memmert, 2011).

Those time-series of formation types together with positions and statistical information allow a wide range of analyses from space- or time-oriented distributions over success in interactions up to tactical aspects. Not least, the condensed game information from above can be used for generating a game protocol, where time, formation type and position data can be completed by semantic information. Such a protocol then builds the basis for the actual game analysis, as it will be discussed in the following.

SOCCER-Based Game Analysis

The basic concept of SOCCER is to handle two types of data: On the one hand syntactic and semantic items taken from video and automatic position recording as well as from expert evaluation, and on the other hand patterns of formations and formation sequences taken from net-based analysis. Recorded data as well as recognized patterns can be anlyzed statistically under the aspects of frequencies and spatio-temporal distributions. The central idea of the approach is that both groups of information can be combined in a compound analysis: The formations and formation sequences build the basis for the understanding of interaction and tactical patterns. They are also useful as a background and/or context for the evaluation of statistical items. In turn, syntactic and semantic items are helpful for understanding and evaluating pattern constructs.

Original data, patterns and results from analyses are organized in a data base. They can be presented in interactive tables, graphics and animations. This allows an arbitrary combination as well as a stepwise resolution of the presented information.

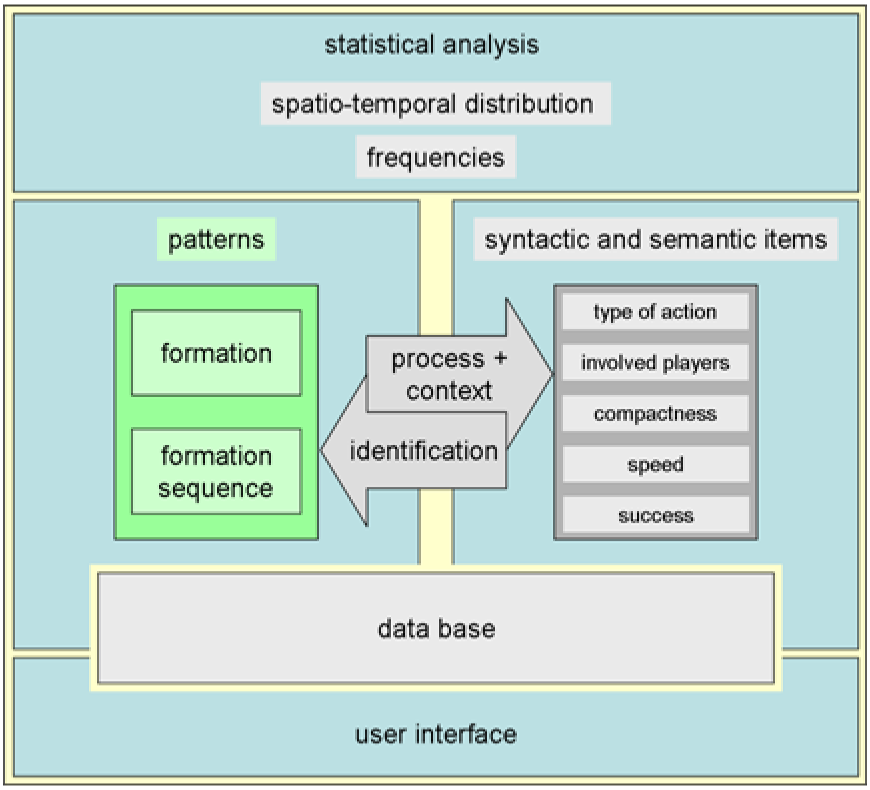

Figure 2. Basic concept of the SOCCER analysis tool (list of items is only a selection).

In the following, three examples of SOCCER-based analyses are presented, followed by a closing section on rule-based semantics analysis:

(1) Distribution analyses. The distribution analysis presented in Figure 3 demonstrates a typical situation: Team A attacks in the formation of type 4, team B reacts with a defense formation of type 3. The distribution matrix shows that this particular interaction happened 523 times (i.e. at 523 seconds) in the corresponding half-time.

In general, the matrix provides the distributions of formations of the teams as well as those of the respective interactions.

Statistical analysis is helpful for a first recognition of normal and of seldom or striking situations. In order to recognize the role they play in the game process, statistical analysis can be combined with animated process analysis.

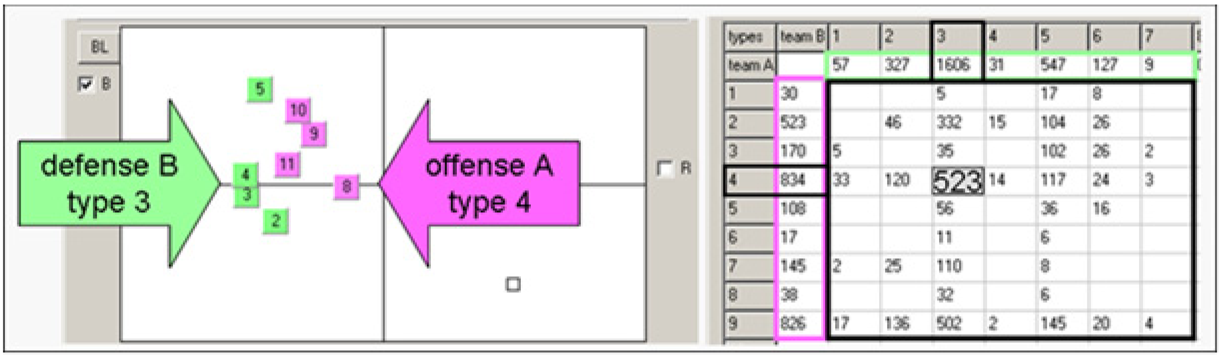

Figure 3. Distribution matrix of formation types and interactions (taken from Perl & Memmert, 2011).

(2) Combined quantitative and qualitative process analysis. As mentioned above, the formation data can be completed by semantic ones like technical or tactical aspects and success. The following example deals with evaluating the success of a team in a given formation interaction. Figure 4 shows from left to right the number of evaluated interactions of a team, followed by the negative ones in absolute numbers and as percentages. Concentrating on the right graphic, it seems that team A has serious problems in the interaction of formation 3 vs. formation 3. However, the absolute numbers are very small, reducing the importance significantly. Also '5 vs. 5' is negative but does not seem very important, whereas '5 vs. 2' seems to be a significant weakness, although the percentage of negative results is only 16. Note, that the presented analyses are only examples, which can be completed arbitrarily if the valuation data is once available.

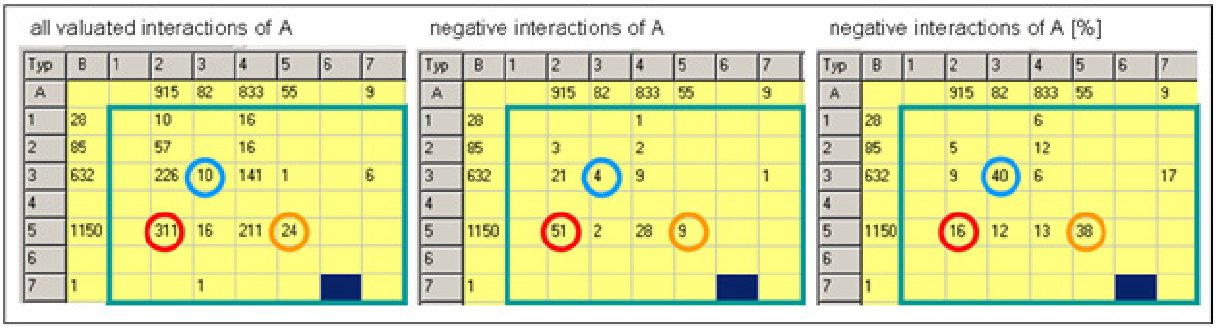

Figure 4. Matrices of valuated team success in the context of formation interaction (taken from Perl & Memmert, 2011).

(3) Net-based tactic analysis. Tactic analysis is done by net-based trajectory analysis. The idea is that at each point in time the formations of tactical groups are identified and can therefore be encoded by a corresponding number and/or color. After training, the network can recognize the

formation and the formation type contained in each data set of the original position data, and is therefore able to map the original process to a trajectory, as it is demonstrated in Figure 5.

The graphic shows a net of neurons, i.e. the small white or colored squares, where each color stands for a formation type like the one in Figure 1. Different neurons of the same color represent variant formations of the same type. Representing those variants by just one characteristic type reduces the number of significantly different items to only about 10, which has two important advantages: On the one hand, it enables statistical analyses on reasonable distributions. On the other hand, the formation trajectories are smoothed and therefore allow easier comparisons between each other. (Note that all specific information is saved and can easily be used for special analyses if needed.)

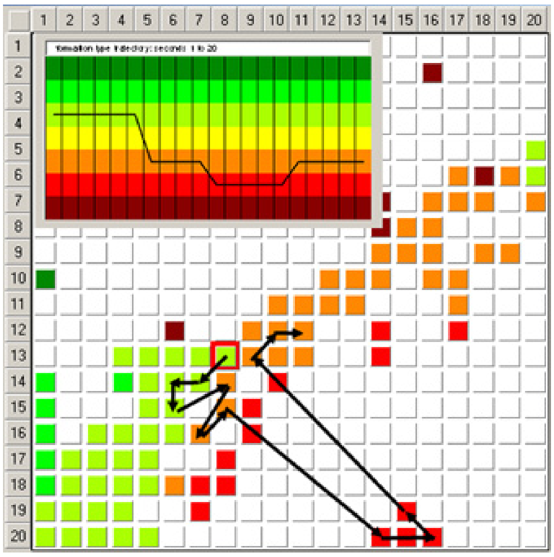

Figure 5. A trajectory of formations on the net and its reduction to a formation type trajectory.

In the presented example from Figure 5 it works as follows: The position data sets of the game process activate corresponding neurons of the network, starting with the one with the red mark. The process then runs through some light green neurons followed by some orange and some red ones and so on. Reduced to the significant types represented by the corresponding colors, the trajectories become much simpler and therefore represent the specific behavior of the corresponding tactical group (see the small embedded graphic on top left).

Such tactical phase patterns can be put in, clustered and finally recognized using neural networks on a second level. In the following, an example of a striking feature analysis is given which demonstrates the way, how a complex analysis of combining quantitative and qualitative aspects works: Figure 6 presents a sequence of corresponding formations of team A and team B. A first glance on the formation phases shows that team B has frequent changes from 2 to 4 and back, while A has analogous changes between 5 and 3. These correspondences can be systematic or arbitrary. A second level analysis may help to answers questions like this and lead to a better understanding of such tactical interactions.

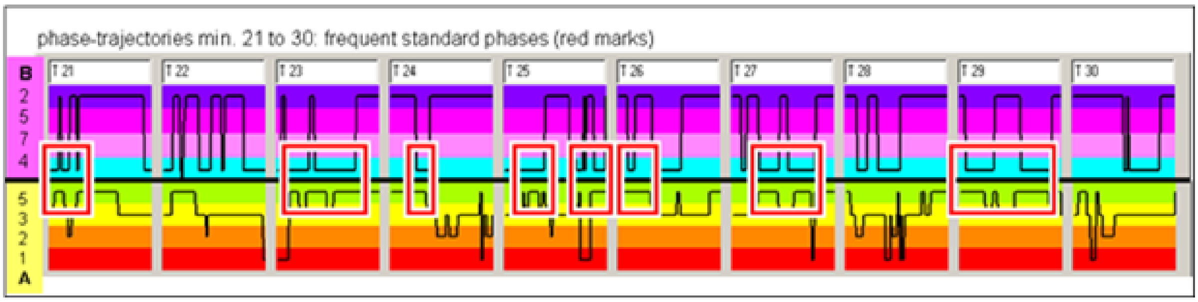

Figure 6. Distribution of a typical pair of formations between minutes 21 and 30.

Rule-Based Semantics Analysis

A last and most challenging task is the one of automatically recognizing semantic information from the game data: Although formations and tactical interactions play an important role for the understanding of the processes in the large scale, the practice of players and coaches demands information on the ball win and loss, ball possessions, starts of attacks and success of offense or defense activities in general. In turn, that particular information on actions improves its value and importance by far, if embedded in the context of the behavior of the involved tactical groups.

The problem, however, is how to get the semantic information from the position information only. The basic idea of one of the authors (Perl) is that the position data of the players and the ball give information about the probability that the player closest to the ball is in contact with it, if the distance between that player and the ball is smaller than a sufficiently small distance. Of course, this assumption is not correct every time, but from the statistical point of view, it helps to deduce the following information: (a) Ball win / loss: The ball contact changes from a player of one team to the other team; (b) Ball possession: Over a certain time interval, players of the same team have ball contact; and (c) Making / receiving a pass: The ball contact changes from a player of a team to a (normally) distant one of the same team. To evaluate single activities, "win", "making", "receiving" and "possessing" can be taken for "successful", while "loss" or "without ball contact" can be taken for "unsuccessful".

Based on this elementary information, processes and their evaluations can be defined. Starting and running an attack, for instance, can start with the ball win, followed by a ball possession, and followed by a pass and so on, normally ending with a ball loss. The final loss, however, naturally does not mean that the whole process was unsuccessful, i.e. only the steps are counted. Finally, positions and contact measuring help to evaluate the tactical behavior of groups or the whole team: One example is the compactness of a team relatively to the ball, which gives important information on the players' positioning and ball orientation.

A second example is the speed of a defense process in getting the ball under control and leading it back to start an attack. A third example is the speed of an offensive process in getting the ball into the opponents half. The SOCCER data preparation and management offers a complex handling of all information about players, their positions, formations, ball contacts and success. This is organized by means of time depending on data vectors which are – based on the integrated data base concept – reflected in interactive tables on the interface. Figure 7 shows how it works.

Every combination of players, groups, formations, ball contact situations, areas on the soccer pitch, and success can be activated just by clicking the regarding input tables, resulting in information about when, where, how successful and how often (absolute, in percentages, compared to the opponent team) the selected situation was. In the example in Figure 7, the analysis focused on the ball possession activities of the offense of team A in the area close to the goal on the right hand side. The result was that formation 3 was 50% dominant in the corresponding points in time t and the formation sequences over the respective last 6 seconds were rather constant (see table on right hand side).

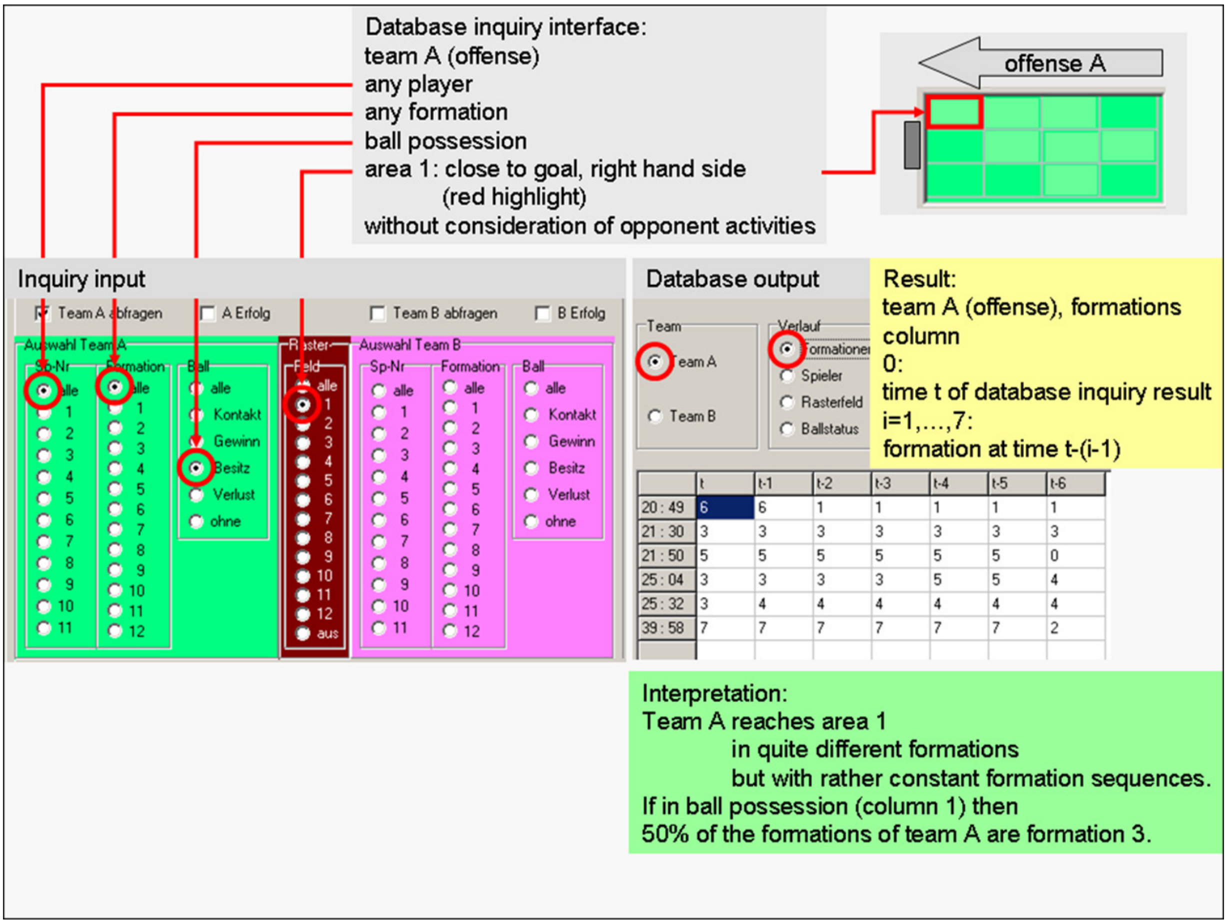

Figure 7. Example of a database inquiry and result interpretation.

Figure 8 demonstrates a specific analysis dealing with ball losses in the context of formation and areas. The result of the presented example is that team A has a symmetric right-left distribution of losses over all formations with a minimum of only 2 losses in the context of formation 4. It should be mentioned, however, that the absolute number of formation 4 is also rather small, i.e. the percentages, which can also be retrieved from the database are an important measure for the ball possession value of formations.

Finally, speed and compactness analyses are special features of SOCCER: The subject to the analysis is the process where the defence stops the opponent attack and passes the ball as far as possible to the own offence in the opponent's half. The process stretches over four points in time, t0, t1, t2, t3, beginning with the ball win and ending with the offence's attack, and is analysed regarding the contexts of formation, tactical groups and specifically involved players. This way, two main questions can be answered:

(1) What is the speed of the process along t0, t1, t2, t3 – i.e. how fast can the team or tactical group react and start counter-attacks?

(2) How does the centroid of the team or tactical group moves related to the ball, and how does its compactness change?

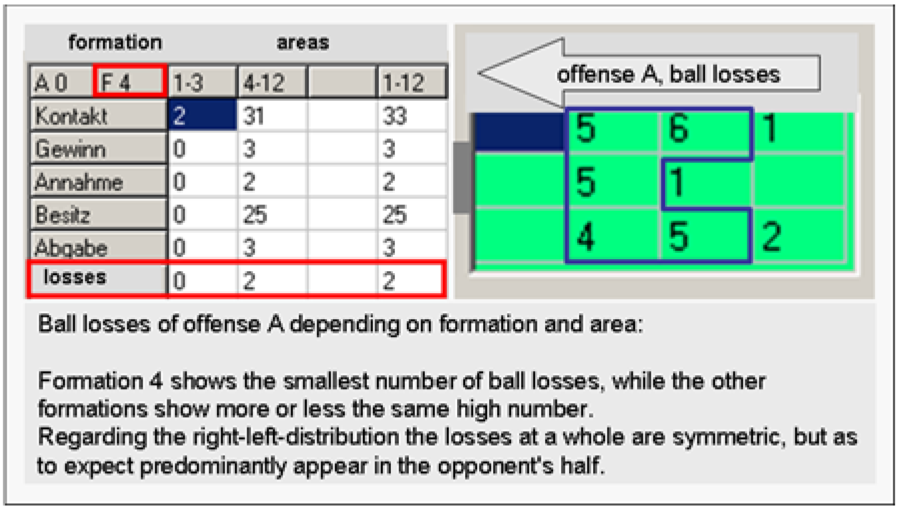

Figure 8. Distribution of ball losses, depending on formation and area.

Experiences from the analysis of more than 40 games shows that only the combination of position-based analysis and semantic analysis in the context of tactical group formations is able to make complex playing processes transparent and understandable, helping to improve tactical behaviour.

Conclusion and Outlook

Over the past years, progress of computer science made it possible to track the players‘ movements and thus to provide position data. Neural network approaches have become a frequently studied and commonly recognized possibility for data analysis and data simulation in sports (Memmert & Perl, 2009b). These studies have demonstrated that additional research questions linked with pattern learning, game analysis and simulation processes can profit by using neural networks. All in all, the role and development of neural networks is now a topic of current discussions in computer sport science and also in performance analysis in soccer.

In the field of game analysis, SOCCER became a tool for automatic analyses and assessments of tactical behavior based on position data for the first time. If validity can be ensured, the planned assessment system will be an important step towards objectification of tactical performance components in team sports. Furthermore, there will be a dramatic speed advantage concerning the evaluation of the position data (from 6-8 hours to 2 minutes). The small effort for data acquisition will enable the accumulation of a vast amount of data and will thus bring new chances for theory construction and practice in soccer.

References

Buschmann, J. & Nopp, S. (2006). *FIFA-World-Cup 2006. Game-Analysis Germany-Argentina.* German Sport University: Cologne.

Buschmann, J. & Nopp, S. (2010). *FIFA-World-Cup 2010. Game-Analysis Germany-England.* German Sport University: Cologne.

Edgecomb, S.J. & Nortona, K.I. (2006). Comparison of global positioning and computer-based tracking systems for measuring player movement distance during Australian Football. *Journal of Science and Medicine in Sport, 9* (1-2), 25–32.

Grunz, A., Endler, S., Memmert, D. & Perl, J. (2011). Netz-gestützte Konstellations-Analyse im Fußball [Net-based constellation analysis in soccer]. *Sportinformatik trifft Sporttechnologie*, 111–115.

Grunz, A., Memmert, D. & Perl, J. (2009). Analysis and Simulation of Actions in Games by Means of Special Self-Organizing Maps. *International Journal of Computer Science in Sport, 8* (1), 22–37.

Grunz, A., Memmert, D. & Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science, 31* (2), 334–343.

Herzog, G. & Retz-Schmidt, G. (1990). Das System SOCCER: Simultane Interpretation und natürlichsprachliche Beschreibung Zeit veränderlicher Szenen [The system SOCCER: simultaneous interpretation and natural language description of time changing scenes]. In J.Perl (Ed.)*, Sport und Informatik,* 95–119. Schorndorf: Hofmann.

Hughes, M & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences, 23* (5), 509–514.

Impellizzeri, F. M., Sassi, A., & Rampinini, E. (2006). Accuracy and reliability of a commercial video-computerised, semiautomatic soccer-match analysis system: Preliminary results. *In Proceedings of the Eleventh Annual Conference of the European College of Sport Science* (pp. 319–326), Lausanne, Switzerland.

Kohonen, T. (2000). *Self-Organizing Maps (3rd. edition)*. Berlin, Heidelberg: Springer.

Mc Garry, T. & Perl, J. (2004). Models of sports contests – Markov processes, dynamical systems and neural networks. In M. Hughes & I. M. Franks (Eds.), *Notational Analysis of Sport* (pp. 227–242). London and New York: Routledge.

Memmert, D. & Perl, J. (2009a). Game Creativity Analysis by Means of Neural Networks. *Journal of Sport Science, 27*, 139–149.

Memmert, D. & Perl, J. (2009b). Analysis and simulation of creativity learning by means of artificial neural networks. *Human Movement Science, 28* (2), 263–282.

Nopp, S. (2012). *Direct vs. Possession Play. Successful team tactic parameters in soccer at national and international elite level.* German Sport University: Cologne.

Perl, J. (2004). A Neural Network approach to movement pattern analysis. *Human Movement Science, 23* (5), 605–620.

Perl, J. (2008). Modelling. In P. Dabnichki & A. Baca (Eds.), *Computers in Sport* (pp. 121–160). Southampton: WIT Press.

Perl, J. & Memmert, D. (2011). Net-Based Game Analysis by Means of the Software Tool SOCCER. *International Journal of Computer Science in Sport, 10* (2), 77–84.

Reep, C. & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society, A, 131,* 581–585.

Stelzer, A. (2004). A new technology for precise local positioning measurement – LPM. In A. Fischer & Vossiek, A., *Microwave Symposium Digest, 2004 IEEE MTT-S International* (pp. 655–658). Linz, Austria.

PART 2

SPECIAL EDITION

Beyond iBeer – Exploring the Potential of Smartphone Sensors for Performance Diagnostics in Sports

Oliver Hummel[1], Ulrich Fehr[2] & Katja Ferger[3]

[1]Software Design and Quality, Karlsruhe Institute of Technology (KIT)

[2] Institute of Sport Science, University of Bayreuth

[3] Institute of Sport Science, University of Gießen

Abstract

In recent years, modern technology significantly improved the possibilities for performance diagnostics in sports. However, the technical equipment used for this purpose is usually still complicated to transport, to set-up as well as to operate and to top it all rather expensive. Off-the-Shelf smartphones on the other hand are relatively cheap, independent of power outlets and familiar to a large range of users. As smartphones nowadays contain sophisticated sensor technology, it seems promising to investigate their potential to cheapen performance diagnostics in sports and thereby make it available to a larger target audience. This article presents the results of an explorative study that demonstrate how a current Android-based smartphone is able to deliver results in the analysis of "bouncing" jumps that are comparable to common force-platforms. Hence it seems worthwhile to create an app able to deliver this service automatically on mobile devices in the near future.

KEYWORDS: PERFORMANCE DIAGNOSTICS, SPORTS, JUMP ANALYSIS, SMARTPHONES, SENSORS

Introduction

Modern technology has been supporting trainers and athletes in performance diagnostics for numerous years (Liebermann et al., 2002; Miah, 2002; Wagner, 2009) and is thus helping to increase success (Haake, 2009) in competition and training as well as to decrease the risk of injury (Zatsiorsky, 2000). Especially the analysis of contact times of diverse jump variations is frequently used as an indicator for the training condition of athletes in disciplines requiring explosive strength (Hennessy & Kilty 2001; Mattes et al., 2010). Approaches to measure such contact times are manifold on the one hand: force platforms (e.g. by Kistler, Switzerland) or systems based on near-ground light-barriers (such as Optojump by Microgate in Italy) are just a few examples. On the other hand, however, all these systems share two central weaknesses. First, they are custom-built and thus expensive and second, they are laborious to set up and operate so that their usage is often limited to financially powerful organizations such as Olympic Centers or larger Clubs and Colleges. Even worse, most of these devices require a power supply and additional measurement equipment like an A/D converter and a recording device (such as a laptop) to operate properly, which constraints their applicability even further.

In order to overcome these constraints, it is obvious that future performance diagnostic devices should be rather based on mobile devices of a reasonable price than on expensive custom-built products. The electronic mass market is indeed offering an interesting, though perhaps not fully obvious solution for this problem: recent mobile phones, so-called smartphones, such as the iPhone or various Android-based models, are lightweight, independent of power outlets, have built-in network connectivity (Bluetooth, WiFi or 3G), and are usually equipped with sensors that make their usage for movement analysis and performance diagnostics a serious possibility. However, beyond various fun apps such as iBeer (cf. http://www.hottrix.com), scientific reports on the use of mobile phone sensors barely exist so far. The few publications that we are aware of, are mainly focusing on innovative applications in the area of assisted living where smartphones have been used for logging the daily activities of subjects (Troiano et al., 2008), or for detecting downfalls of elderly people (Sposaro & Tyson, 2009). The only publication proposing the usage of the sensors in mobile phones in a sports context is the work of McNab et al. (2011) who propose to use an iPhone's acceleration sensors for recording the throws of cricket players. Further recent works have also discovered the mentioned advantages of smartphones though, but have been merely using them as a base unit for the recording of motion data captured with dedicated sensors (cf. e.g. Strohrmann et al., 2012).

Nevertheless, recent smartphones seem to impose themselves as a simple and affordable alternative to traditional performance diagnostics systems that are easily usable on an daily basis in the field so that a clear improvement in training quality could be expected (cf. e.g. Bauersfeld & Voss, 1992). However, it is not yet clear whether their built-in sensors will be able to fulfil such high hopes? After all, smartphones are still mainly intended as fun product or perhaps as “mobile offices” for business people, but not as serious measurement equipment in motion analysis. Unfortunately, phone manufacturers are thus not routinely publishing the technical performance data of the sensors they use so that their real potential for performance diagnostics or motion analysis in sports still needs to be investigated.

The first contribution of our article is thus a coarse comparison of sampling rate and measurement range of various current smartphone models. Furthermore, we present and discuss the result of a series of calibration jumps used to determine the acceleration pattern of two-legged bouncing jumps in comparison with a force platform. Based on the characteristics found for these jumps we have carried out a validation experiment with a group of six physical education students showing promising results for a potential development of an automated smartphone jump analysis app. Considering the facts of the experience collected from these jumps, we also present some ongoing work and preliminary results from automating the recognition of such jumps on the smartphone. We finally round off our article with a summary of our findings.

Background

In numerous sports, the jumping power is a performance-determining parameter and thus often represents a performance-limiting factor at the same time. Hence, in the context of training management, discipline specific jumping performance is crucial. In order to review and analyse the jumping performance typically three jump types are used: counter movement jump, squat jump and drop jump (Wank & Heger, 2009). Multiple jumps with flexing only in the ankle joint (sometimes also called bunny or bouncing jumps) also provide valuable information on the onset of fatigue and are therefore often used for the determination of strength-endurance (e.g. by a Bosco test – Bosco et al., 1983). Since the latter allow carrying out a large number of jumps with similar characteristics and relatively little disturbing movement effects, they form a particularly interesting target for the analysis of novel measurement equipment and hence have been used in the explorative study we present in this article.

Currently, force platforms and contact mats are the primary means used for diagnosis and analysis of sport specific jumping power in order to capture both dynamic and kinematic parameters as accurately as possible. Parameters such as maximum force and maximum power increase can be determined by means of the force-time curves (Frick, Schmidtbleicher & Wörn, 1991). In doing so, the current training skill of an athlete can be detected. However, it is extremely important to measure, document and analyse the performance history systematically on a regular basis. As mentioned before, the available measurement equipment is rather stationary, requires accuracy in use and often a dedicated analysis of delivered data so that it significantly complicates training planning and management. At this point we aim on developing a simple smartphone app for providing relevant information and feedback on performance data quickly and efficiently during a training session.

Sensor Data Capturing on Android Devices

According to Google (2012), Android devices may provide various built-in sensors for collecting information on motion, environment and position of a device. This does not only include GPS sensors that have been regularly used in the sports and outdoor industry for more than a decade (see e.g. Bouten et al., 1994), but also orientation as well as acceleration sensors and sometimes even gyroscopes. Perl et al. (2011) have summarized various uses of these sensor types (independent of mobile phones) and respective analysis techniques for sports science in an recent overview article. In order to obtain such data from mobile phone sensors, the Android operating system is offering a simple API for determining which sensors are available on a device as well as for finally retrieving the raw data.

The mobile phone we have used for the experiments described in this article is an Xperia Ray produced by Sony Ericsson (2011) that costs about 200 Euros. Compared to other recent phones of competitors, the Xperia Ray, containing a 1 GHz single-core Snapdragon CPU, is relatively small and lightweight, as it measures merely 111 x 53 x 9.4 mm and weighs about 100 g. For our experiments we were mainly interested in the built-in three-axis accelerometer of the phone, although it also offers a three-axis orientation sensor as well as a GPS. As described in the next subsection, sensor values can be captured via a simple Java program for the Android operating system used in the phone, which we have updated to version 4.0.3.

Software

Programmatically, reading the sensor data on an Android phone is a relatively simple undertaking as the following two Java snippets illustrate. It requires only two basic actions, namely registering a listener for the desired sensor type and processing sensor events whenever the listener is called back. As visible in the first snippet, it is necessary to select the desired sensor type and the desired sensor delay (we chose the “fastest” for obvious reasons). Unfortunately, it is not possible to directly select a desired sampling rate, most likely due to the numerous different types of phones and tablets that need to be supported by the Android operating system. Therefore, it merely used to be possible to choose from four so-called sensor delay constants (FASTEST = 0 ms, GAME = 20 ms, UI = 60 ms und NORMAL = 200 ms) before Android 3.0. Since then it is also possible to specify the sensor delay directly as a numerical value, however, to quote the Android documentation, “The delay that you specify is only a suggested delay (… and) There is no public method for determining the rate at which the sensor framework is sending sensor events to your application". In other words, every device delivers sensor events only according to its capabilities and depending on the actual change rate of the physical sensor.

The following code snippet distills the four lines of code that are required to select and initialize a sensor (the accelerometer in this case) in an Android app.

```
// initialize motion detection

sensorDelay = SensorManager.SENSOR_DELAY_FASTEST;

sensorMgr = (SensorManager) getSystemService(SENSOR_SERVICE);

Sensor sensor =
       sensorMgr.getDefaultSensor(Sensor.TYPE_ACCELEROMETER);

sensorMgr.registerListener(this, sensor, sensorDelay);
```

Once a SensorListener is registered like this, the listener method shown in the next snippet is called back whenever a sensor event, i.e. a change in the sensor values, occurs. In other words, a sensor usually creates sensor events only when its measured values have changed, which in turn implies that a phone moved with quickly changing motions, such as through shaking it, should deliver more sampling data per time unit than a phone in a static position. The snippet also illustrates how acceleration values for the three axes and a timestamp can be extracted from the submitted SensorEvent data object.

```
// process sensor event

public void onSensorChanged(SensorEvent se) {

      long time = se.timestamp;

      float ax = se.values[0];

      float ay = se.values[1];

      float az = se.values[2];

      ...

}
```

Once the acceleration values have been obtained as described it is easy to calculate the absolute acceleration (that ignores the phone's concrete orientation) from this vector through taking the root of the three squared values, as follows:

```
      float aa = FloatMath.sqrt(x*x + y*y + z*z);
```

For analysis purposes it also makes sense to store captured data on the phone's SD card or internal memory from where it could be transferred via USB or Bluetooth to any arbitrary PC or laptop.

AccelLogger App

Based on the above data collection code, we have created a simple app that is able to measure acceleration values and to store them as CSV files on the phone's permanent memory. We called the app AccelLogger and have made it freely available in Google's PlayStore (https://play.google.com/store/apps/details?id=net.zehnkampf.accellogger) in order to allow an easy reproduction of our experiments. Beyond the plain recording of data values it is also able to display them as acceleration-time curve on the phone as shown in the following screenshot of seven "bunny jumps". It also gives a first impression of the characteristics of these jumps that we will analyze in more detail later.

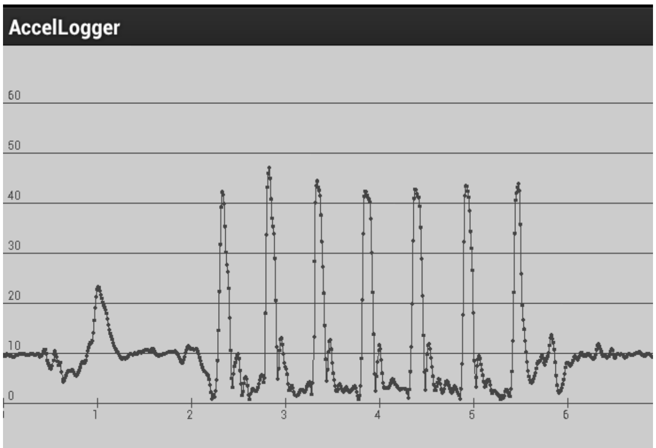

Figure 1. Screenshot of seven "bunny jumps" taken from the AccelLogger smartphone app.

As this diagram view allows an immediate visual examination of the recorded values, it can also be used as a first plausibility check to uncover problems during data capturing. Moreover, the AccelLogger app is able to fundamentally analyze the recorded data in order to determine the actually achieved sampling rate as well as the measuring range for all three axes as shown in the following screenshot.

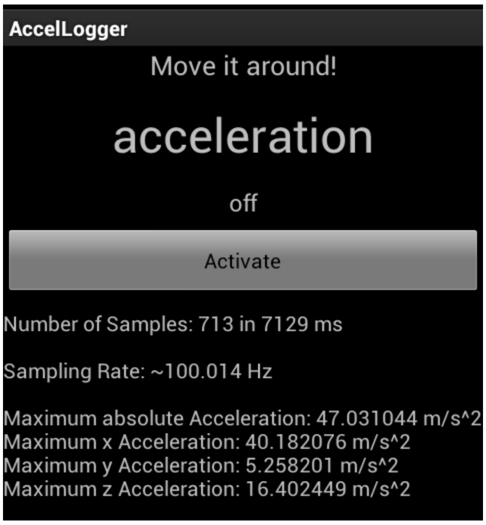

Figure 2. Screenshot of AccelLogger's main view after a series of jumps.

Finally, the app also allows selecting one of the four sensor delay constants mentioned before so that it is easy to determine the achievable sampling rates for each setting as we will present in the following section.

Sensor Testing and Calibration

In order to avoid influencing our measurements through background tasks on the phone and to influence the phone's processor as little as possible during data recording, we ended all available tasks and put the phone into flight mode prior to executing the measuring experiments. Furthermore, we did not do any additional data processing during the recording; we simply stored the data as received from the sensor on the phone's SD card as described before. The subsequent data analysis was carried out with Microsoft Excel on a Windows laptop to which we transferred the files later. Initially, we were most interested in finding the effective sampling rates possible with the Sony Xperia Ray based on the predefined sensor delay constants of the Android operating system mentioned before. For that purpose, we carried out two simple experiments, first we recorded the acceleration values delivered by the phone while it was lying still on the ground for 30 seconds and second we recorded them while a subject was jumping around with it for around 30 seconds. The following table presents the approximate sampling rates obtained.

Table 1. Overview of sampling rates achieved with various sensor delays on the Xperia Ray.

Movement	FASTEST	GAME	UI	NORMAL
None	86.7 Hz	46.6 Hz	13.3 Hz	13.3 Hz
Jumping	100.0 Hz	50.0 Hz	14.4 Hz	14.4 Hz

Although the values fluctuate slightly in repeated runs, probably due to noise and light building vibrations, we spared ourselves a more detailed analysis, since this experiment was merely intended to basically show that the Xperia Ray is capable of capturing acceleration data with a useful sampling rate of 100 Hz. As also visible in Table 1, the measured values for NORMAL and UI sensor delay are identical for some reason so that the Xperia Ray is obviously not able to support sensor delays larger than roughly 60 milliseconds.

Phone Comparison

We have been able to briefly run our acceleration logger app on two other Android-based phone types so far (Samsung Galaxy S II (Samsung, 2010) and Motorola Razor) and found surprisingly large differences in their measuring ranges as detailed in the following table.

Table 2. Comparison of sensor performance of three current Android phone models contrasted with iPhone 4 data taken from the literature.

Phone Type	Highest Sampling Rate	Measuring Range per Axis	Maximum Absolute Acceleration
Sony Xperia Ray	100 Hz	± 40 m/s^2	~ 70 m/s^2
Samsung Galaxy S II	100 Hz	± 20 m/s^2	~ 35 m/s^2
Motorola Razor	100 Hz	± 80 m/s^2	~ 139 m/s^2
iPhone 4[1]	*unknown*	± *20* m/s^2	~ *35* m/s^2

[1] As reported by McNab et al. (2011).

We have obtained the reported values by shaking the phone as hard as we could for a few seconds and analysing the measured values afterwards. The Xperia Ray and the Razor have shown a good performance during these tests and thus should both be readily usable for motion analysis in sports. The measuring range of the Galaxy S II, however, seems to be too small to analyse jumps in a useful way as its compressed acceleration-time curves are not expressive enough. This brings us to the conclusion that motion analysis with smartphones is currently highly dependent on the product used. This assumption is also backed up by a personal conversation with one of the authors of Strohrmann et al. (2012) who confirmed that their attempts to use a Sony Xperia Active for this purpose failed due to the limited measurement range so that they were forced using dedicated Shimmer sensors for their experiments.

The following remainder of this section is describing our efforts to obtain a basic understanding for the capabilities of the Xperia Ray in performance diagnostics. We describe the analysis of some initial jumps on a force platform that we performed in order to understand the acceleration-time curves captured by the phone and for being able to identify the jumps in the data delivered by the phone.

Understanding Sensor Values

In this subsection we explain how we compared the acceleration-time curves of the Xperia Ray with the force-time curves of the force platform we have used for our initial experiments. The latter was a model by Kistler (Switzerland) operated at 125 Hz with a threshold of 10 N. As shown in the following photograph, the smartphone was placed in a cell phone case that was attached to the lower back of our subjects.

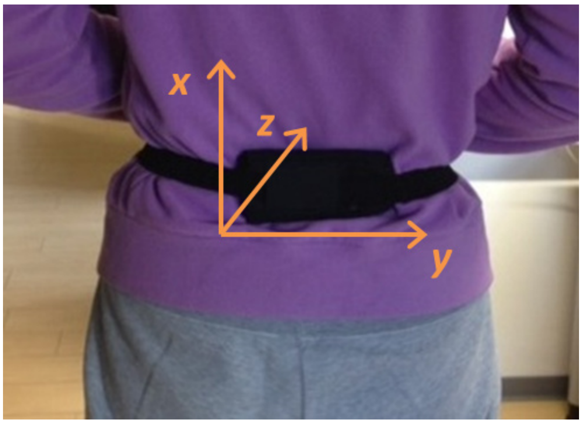

Figure 3. The Xperia Ray as attached to the subjects during our experiments.

For illustration purposes, the axes of the phone's accelerometer have been added to the figure in order to simplify the understanding of the following figure, showing an excerpt (basically the landing of one jump) from the recording of a series of bouncing jumps.

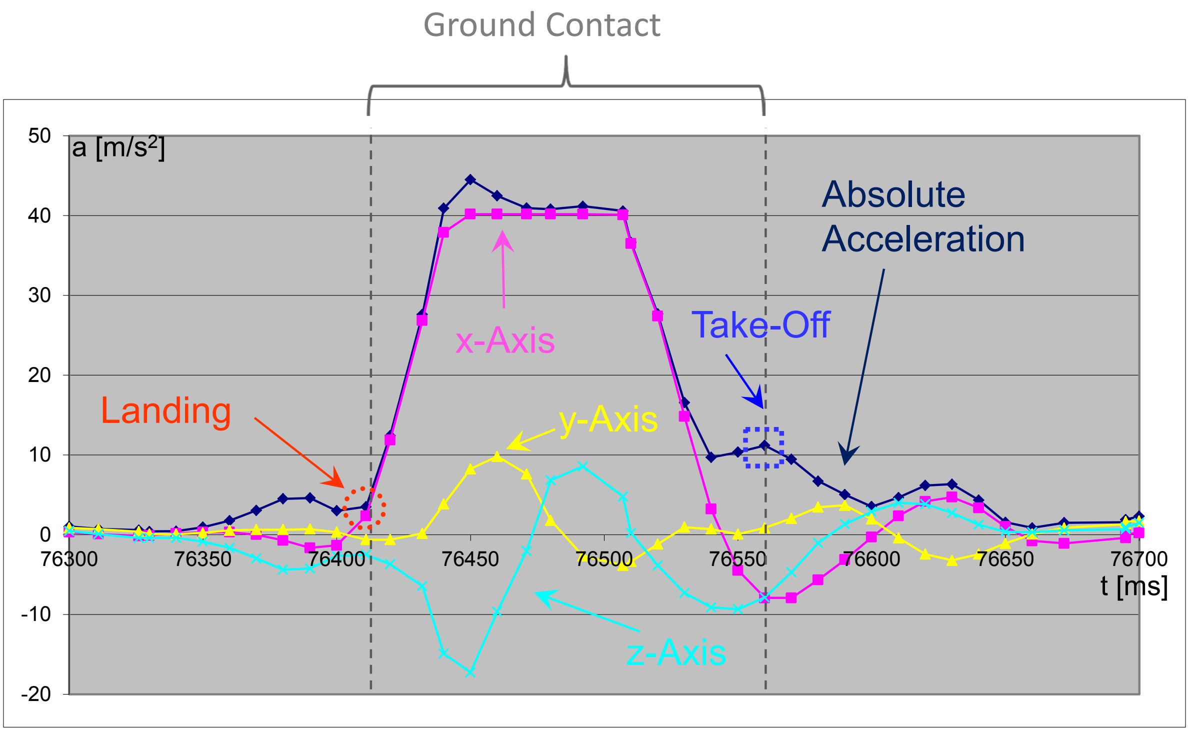

Figure 4. Analysis of one bunny jump as measured on the smartphone.

Figure 4 illustrates how the curve for the absolute acceleration is composed from the acceleration vectors for the three axes, as described before, through taking the root of the squared sum of the acceleration vector. As visible, the x-axis is depicting the up and down movement of the jumper, thus having the biggest influence on the absolute acceleration in this case. The y-axis captures potential sidewise movements while the z-axis represents the forward or backward movements caused by slight hip movements during the jump.

Through our analysis and comparisons with the force platform data we figured that take-off and landing times can be determined with reasonable reliability in the absolute acceleration as follows – a take-off can be recognized as the small local peak (caused by the superimposition of the three axes) after the large landing acceleration; the first ground contact of the landing can be determined as the last value before the significant increase in acceleration caused by the landing.

Figure 5 exemplarily contrasts two jumps and the two measurement approaches with each other. As to be expected, the force platform delivers smooth force-time curves where the beginning and the end of the jumper's ground contact is detectable in the upper diagram. Fortunately, as also visible in the lower diagram of the figure, the curve of the smartphone's absolute acceleration as described before also turned out to be highly characteristic.

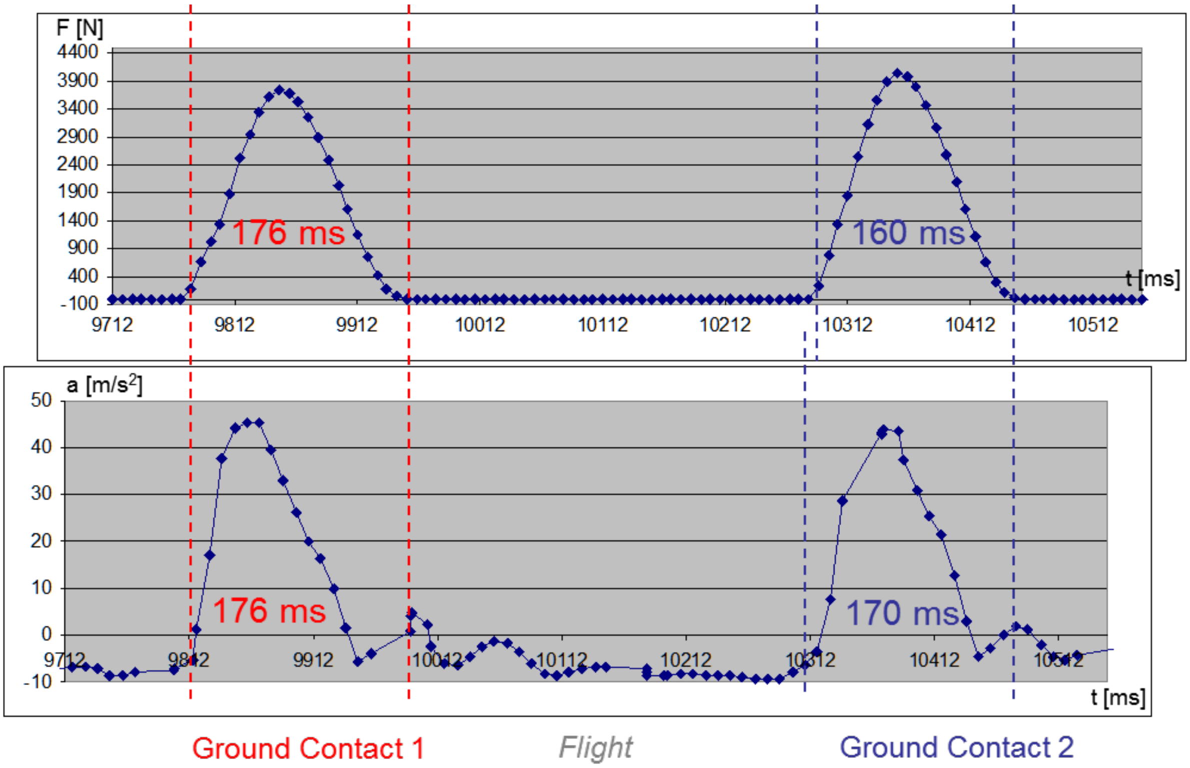

Figure 5. Comparison of two jumps measured on the force platform (upper diagram) and the smartphone (lower diagram). Offset caused by slight delay until impulses reach the smartphone.

Assuming that every value above the threshold of 10 N on the force platform is a ground contact of the jumper, the analysis yields two ground contacts of 176 ms and 160 ms. The phone sensor looks very similar although it is roughly 20 ms delayed in comparison with the force platform. However, this can be easily explained with the delay that is needed until the ground contact is "propagated" through the jumper's body. Fortunately, for detecting the absolute ground contact and flying times in the field this minimal shift is irrelevant anyway. Please note that we have subtracted earth gravity from the acceleration values in this figure in order to better illustrate the flight phase between two ground contacts. Applying the detection rules for the smartphone as defined before yields one ground contact of also 176 ms and one of 170 ms. The deviation of 10 ms to the reference value of the force platform seems to be mainly caused by an unfortunate occurrence of sampling events in this case.

Evaluation

To verify the accuracy of the Xperia Ray (Sony Ericson, 2010) for measuring flight- and ground contact times in a more generalized setting, the data of 48 two-feet bunny jumps from 6 male physical education students was recorded on the smartphone that was placed in a tight-fitting softcase and fixed with a belt at the hip as shown in Figure 4. The collected data was compared with data recorded simultaneously by a force platform (Kistler, Switzerland, model 9253A11), this time operated at 1,000 Hz with a threshold of 25 N. Based on the detection rules presented in the last subsection, we derived ground contact and flight times in milliseconds manually from the acceleration-time curves of the mobile phone (with the help of Microsoft Excel diagrams), since the data of the force platform was easier to interpret, a small VBA script simplified this task for us.

The test instruction for the subjects was: "Starting from a standing position, feet slightly apart, please jump off with both feet, during the flight prepare for a landing with a short ground contact and a quick rebound. Leave your hands on the hips – no arm swing is allowed to support the jumps!". Thus, the subjects started from a standing position on the force platform. We evaluated eight continuous two-feet ankle hops as described from every subject.

Results

The following table summarizes the flight times we have measured for our subjects and the cumulated deviations between the force platform's reference values and the smartphone measurements.

Table 3. Summarized flight times from our evaluation experiments.

Subject	Flight Times	1	2	3	4	5	6	7	8	Cumulated Deviation
1	Smartphone	310	310	309	340	310	270	260	260	
	Force Platform	325	309	320	343	310	271	253	261	39
2	Smartphone	70	110	90	90	100	120	110	100	
	Force Platform	78	97	87	81	84	109	100	101	71
3	Smartphone	159	141	110	110	100	90	90	110	
	Force Platform	160	141	105	103	98	89	76	101	39
4	Smartphone	170	170	180	180	170	180	170	200	
	Force Platform	160	179	184	173	161	165	169	180	75
5	Smartphone	140	100	100	90	90	90	100	110	
	Force Platform	124	94	84	88	86	80	88	81	95
6	Smartphone	100	140	110	130	130	110	120	110	
	Force Platform	101	128	126	135	136	118	135	119	72

Based on the numbers presented in Table 3, the average deviation over all 48 jumps is 8.15 ms with a standard deviation of 6.22 ms. Based on the average flight time measured with the force platform, which is 147.83 ms, this yields a relative error of about 5.5 percent. The maximum deviation from the force platform's reference is 29 ms in the last jump of subject five; all other values are not larger than 20 ms. It also seems interesting to mention that the first two subjects reached extremely different flight times although their ground contact times presented in the following table are very similar.

Overall, the results measured for the ground contact times of the subjects with the help of the smartphone look also quite encouraging; however, subject number 5 stands out with a deviation from the force platforms reference values that is almost twice as large as the second highest deviation.

Table 4. Summarized contact times from our evaluation experiments.

Subject	Contact Times	1	2	3	4	5	6	7	8	Cumulated Deviation
1	Smartphone	190	170	181	160	140	150	150	149	
	Force Platform	182	162	177	151	148	152	150	159	49
2	Smartphone	160	170	160	160	160	160	150	170	
	Force Platform	179	174	169	182	167	178	168	175	102
3	Smartphone	161	160	150	140	139	140	150	140	
	Force Platform	167	155	159	145	144	143	158	154	55
4	Smartphone	140	150	130	140	130	140	130	120	
	Force Platform	140	142	142	145	143	146	139	141	74
5	Smartphone	170	160	160	170	170	170	160	160	
	Force Platform	192	185	190	187	185	195	188	193	195
6	Smartphone	170	170	170	180	180	180	190	170	
	Force Platform	164	158	172	165	175	172	174	157	77

The average deviation of the ground contact times calculates to 11.5 ms with a standard deviation of 8.09 ms, the relative error based on the average ground contact time of the force platform, which is 164.3 ms, is 6.9 percent. The largest individual deviation here is again from subject number five (33 ms in his last jump) with maximum values around 20 ms for the other subjects.

Discussion

Overall, the results of this first larger evaluation are quite encouraging: Except for one series of values in the ground contact times, practically all other measured values are in the range of the theoretically possible measurement accuracy of the smartphone. An inbuilt inaccuracy of up to 9 milliseconds results from its sampling rate of 100 Hz, namely if the mobile phone just cannot measure an acceleration at the time of $t = 0$ that happens exactly one millisecond later (at $t = 1$) and is detected by the force platform at this very moment. In this case, the smartphone will detect this acceleration only during the next sampling sequence 9 milliseconds later. Another unpleasant effect detected is the slight deviation in the timestamps of the smartphone's sampling data that are not always 10 milliseconds apart of each other (this is also the reason why we rounded the timestamps originally delivered in nanoseconds to milliseconds). However, not only the capabilities of the smartphone's sensors influence the validity of our study, but the measurement of the force platform as well. Also it was operated at 1000 Hz, the threshold of 25 N is quite high and an analysis of the raw data reveals some delay time caused by the inertia of the mechanics in these measurements as well. Hence, to obtain even better ground truth data for calibrating the smartphone data, it seems helpful to analyse a series of jumps with high speed video to detect the exact moment of ground contact and final take-off in the future. Furthermore, pre-processing the smartphone data in order to achieve equidistant sampling times should also improve data quality.

This explorative study was also interesting from another perspective as it yielded two more important results. The first one is that it is indeed possible to analyse bouncing jumps of different subjects with the same method as the absolute acceleration-time curves of their jumps look very similar and have the same characteristic inflection points. However, it also demonstrated that there are still subtle cases in which the analysis of the absolute acceleration is not sufficient. From this perspective, subject five is of particular interest– while we found a few jumps where the inflection points in the absolute acceleration curve started to blur, it was most often the case with this subject. Analysing the acceleration axes of his measurements separately revealed a relatively large movement in the z-axis. In other words, he was moving his hip forward and backward during the jumps significantly more than all other subjects so that the absolute acceleration was influenced negatively. Based on this insight, we believe the first step for improving our approach should be the individual analysis of the three acceleration axes (especially the x-axis) since this might deliver better results for an upcoming automated jump analysis app. Moreover, additional movements as detected for this subject might become a good indicator for judging the proper execution of a jump, which in turn could become an additional indicator for an increasing fatigue of an athlete.

Outlook

Although the results presented so far are very promising, it is obviously necessary to automate the recognition of relevant parameters (i.e. at least flight and contact times) on the phone in order to provide a practically usable training support system. Hence, we have started some work in this direction which we want to describe in this section. At a first glance, the automation of jump detection seems to be a relatively simple undertaking that can be based on the passing of simple threshold values. However, the measurement ranges of different phones and the cleanliness of the jumps are crucial aspects in this context that unfortunately make automated detection more complicated. It gets even worse, as not only thresholds, but also prior maxima need to be recognized automatically for properly detecting the correct phase of a jump, which is also influenced by the sampling rate of the used device. Hence, we are currently only working on a prototypical automation of this challenge through a simple algorithm that is optimized for the measurement range and sampling rate of the Xperia Ray and not yet generalized for other phone types. Obviously, the next goal should be to integrate a generalized version of this algorithm into the AccelLogger app so that it is able to present analysis results directly after a series of jumps was executed.

The detection algorithm we have implemented so far is relatively primitive as it is aiming on the detection of the inflection points described in the context of Figure 4 by a simple comparison with previous and subsequent values. Moreover, we have implemented some sanity checks intended to guarantee that the inflection points are found in a meaningful range of acceleration values only. In other words, not every movement of the phone should be detected as a jump. A version of the AccelLoger app enhanced with this algorithm is able to present the detected values in a simple tabular view and also marks them in the acceleration-time diagram with blue (take-off) and red dots (landing) as shown in the following screenshot.

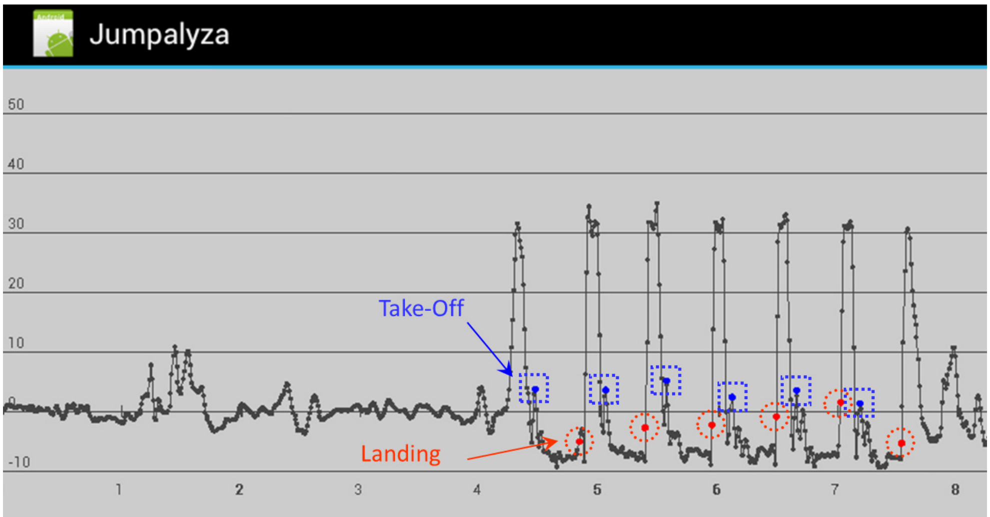

Figure 6. Graphical representation of recognized jumps in the Jumpalyza app.

The algorithm works reasonably fast on the Xperia Ray so that no delay is recognizable for series of around ten jumps and even longer series can easily be analyzed in under one second. However, as indicated before, the algorithm is still adapted on the sampling rate and the measuring range of the Xperia Ray and thus it is not yet possible to execute it on other smartphone types with different sensor characteristics without adaptations.

Conclusion

In this article, we have investigated the applicability of modern off-the-shelf smartphones for performance diagnostics in sports. This idea has been motivated by the high price and relatively complex handling of existing systems for this purpose (such as force platforms). Moreover, most existing systems are not mobile in the sense that they would easily fit into a small bag or could be used in the field without continuous power supply. Hence we have identified multiple two-legged jumps with flexing in the ankle points only (also known as bouncing or bunny jumps) and their regular movement pattern as the ideal candidates to investigate the potential of modern smartphones in jump analysis. We have used an Android based Sony Xperia Ray (Sony Ericsson, 2011) to capture acceleration-time curves of such jumps and analyzed them in comparison to force-time curves as delivered by common force platforms.

The main findings of our work – that goes far beyond the only other publication in this direction that we are aware of (McNab et al., 2011) – are as follows. First, we have found that the Xperia Ray (as well as a Motorola Razor) provides acceleration sensors with a reasonable – though still improvable – sampling rate of 100 Hz and measuring range (± 40 m/s^2 resp. ± 80 m/s^2) that make them interesting for more serious use cases than just the iBeer fun app, for instance. Moreover, we have presented an explanation of the acceleration-time curves measured by smartphone sensors in the context of bouncing jumps and identified clear indicators for take-offs and landing. In order to evaluate the generalizability of this approach we have tested a simple data recording app with six student subjects in comparison to a force platform. This experiment has revealed that even the absolute acceleration values of the smartphone sensors are able to detect ground contact and flight times of bouncing jumps with precision similar to a force platform. The relative error is already below seven percent and even the average absolute error from this investigation is with 11.5 ms near the range determined by the sampling rate of the current smartphone generation (± 10 ms). Based on some outlying results detected for one subject, our study has also revealed various starting points (such as analyzing the three individual accelerometer axes) that might help increasing the precision and the generalizability of our approach in the future.

Based on our preliminary experience with implementing an automated recognition of flight and contact times on the Xperia Ray, we are convinced that a full automated smartphone based jump analysis system should be possible in the not too distant future. Since temporary results from the analysis of drop jumps also delivered promising results, we envision the development of an Android app that is able to automatically analyze various types of jumps in the field in real-time. Furthermore it also seems feasible that even the ground contact times of take-offs in long, high or triple jump or even in sprints can be recognized with such an app in the medium term future. The availability of such a relatively cheap and ultra-mobile performance diagnostics system would certainly make this important technique of training science available to a much larger range of trainers and athletes as this is the case today.

References

Bauersfeld, M. & Voss, G. (1992). *Neue Wege im Schnelligkeitstraining [New ways in speed training]*. Münster: Philippka.

Bosco, C., Luhtanen, P., & Komi, P. (1983). A simple method for measurement of mechanical power in jumping. *European Journal of Applied Physiology and Occupational Physiology*, *50* (2), 273–282.

Bouten, C., Westerterp, K., Verduin, M. & Janssen, J. (1994). Assessment of energy expenditure for physical activity using a triaxial accelerometer. *Medicine and Science in Sports and Exercise, 26* (12), 1516–1523.

Frick, U., Schmidtbleicher, D. & Wörn, C. (1991). Vergleich biomechanischer Messverfahren zur Bestimmung der Sprunghöhe bei Vertikalsprüngen [Comparison of biomechanical measurement methods for the estimation of the jumping height at vertical jumps]. *Leistungssport, 21* (2), 48–53.

Google (2012). Android API Guides: Sensors Overview. Retrieved November 25, 2012 from http://developer.android.com/guide/topics/sensors/sensors_overview.html

Haake, S.J. (2009). The impact of technology on sporting performance in Olympic sports. *Journal of Sports Sciences*, *27* (13), 1421–1431.

Hennessy, L., & Kilty, J. (2001). Relationship of the stretch-shortening cycle to sprint performance in trained female athletes. *Journal of Strength & Conditioning Research*, *15* (3), 326–331.

Hongman, W., Xiaocheng, Z. & Jiangbo, C. (2011). Acceleration and Orientation Multisensor Pedometer Application Design and Implementation on the Android Platform. *Proceedings of the First International Conference on Instrumentation, Measurement, Computer, Communication and Control* (pp. 249–253), Washington, USA.

Kwapisz, J., Weiss, G. & Moore, S. (2011). Activity recognition using cell phone accelerometers, *ACM SIGKDD Explorations Newsletter, 12* (2), 74–82.

Liebermann, D.G., Katz, L., Hughes, M.D., Bartlett, R.M., McClements, J. & Franks, I.M. (2002). Advances in the application of information technology to sport performance. *Journal of Sports Sciences*, *20* (10), 755–769.

Mattes, K., Habermann, N., Korn, M. & Mühlbach, T. (2010). Trainingsbegleitende Analyse und Steuerung von Schrittmerkmalen bei Sprints und Sprungläufen im

Nachwuchstraining. BISp-Jahrbuch: Forschungsförderung 2009, 211–218. Retrieved November, 30, 2012, from http://www.bisp.de/cln_090/nn_113306/SharedDocs/Downloads/Publikationen/Jahrbuch/Jb__200910__Artikel/Mattes__211__218,templateId=raw,property=publicationFile.pdf/Mattes_211_218.pdf

Miah, A. (2002). *Sport technology. History philosophy and policy*. Amsterdam: JAI.

McNab, T., James, D. & Rowlands, D. (2011). iPhone sensor platforms: Applications to sports monitoring. *Procedia Engineering*, *13*, 507–512.

Perl, J., Memmert, D., Baca, A., Endler, S., Grunz, A., Rebel, M. & Schmidt, A. (2011). Sensors, Monitoring and Model-Based Data Analysis in Sports, Exercise and Rehabilitation. In: D. Huei Lai, R. Begg, M. Palaniswami (Eds.), *Healthcare Sensor Networks: Challenges Toward Practical Implementation* (pp. 375–405), CRC Press.

Samsung (2010). Samsung Galaxy S II, Retrieved October 5, 2012 from http://www.samsung.de/de/consumer/detail_printerfriendly.aspx?apath=Privatkunden/Mobil/Mobiltelefone/MultimediaInfotainment/samsunggalaxysii/GT-I9100LKADBT

Sony Ericsson (2011). XPERIA Ray Whitepaper. Retrieved June 17, 2012 from http://www-support-downloads.sonymobile.com/st18/whitepaper_EN_xperia_ray.pdf

Sposaro, F. & Tyson, G. (2009). iFall: An android application for fall monitoring and response. *Proceedings of the International Conference on Engineering in Medicine and Biology Society*.

Strohrmann, C., Gravenhorst, F., Latkovic, S. & Tröster, G. (2012). Development of an Android application to estimate a runner's mechanical work in real time using wearable technology. *Proceedings of Sportinformatik 2012.*

Troiano, R., Berrigan, D., Dodd, K., Mässe, L. Tilert, T. & McDowell, M. (2008). Physical Activity in the United States Measured by Accelerometer. *Journal of Medicine & Science in Sports & Exercise*, *40* (1)*,* 181–188.

Wagner, K. (2009). Entwicklung von Mess- und Informationssystemen (MIS) für die Trainingssteuerung in den Jahren 2008-2009. *Zeitschrift für angewandte Trainings-wissenschaft*, *16* (2), 15–22.

Wank, V. & Heger, H. (2009). Sprünge. In A. Gollhofer & E. Müller (Eds.), *Handbuch Sportbiomechanik [Handbook sports biomechanics]*. Schorndorf: Hofmann.

Zatsiorsky, V.M. (2000). *Biomechanics in sport. Performance enhancement and injury prevention.* (Volume IX of the Encyclopaedia of sports medicine). Oxford, Malden: Blackwell Science.

Heart Rate Variability During Physical Exercise

Heike Leutheuser & Bjoern M. Eskofier

Digital Sports Group, Pattern Recognition Lab, Department of Computer Science, University Erlangen-Nuremberg, Germany

Abstract

The normal oscillation of the heart rate is called Heart Rate Variability (HRV). HRV parameters change under different conditions like rest, physical exercise, mental stress, and body posture changes. However, results how HRV parameters adapt to physical exercise have been inconsistent. This study investigated how different HRV parameters changed during one hour of running. We used datasets of 295 athletes where each dataset had a total length of about 65 minutes. Data was divided in segments of five minutes and three HRV parameters and one kinematic parameter were calculated for each segment. We applied two different analysis of variance (ANOVA) models to analyze the differences in the means of each segment for every parameter. The two ANOVA models were univariate ANOVA with repeated measures and multivariate ANOVA with repeated measures. The obligatory post-hoc procedure consisted of multiple dependent t tests with Bonferroni correction. We investigated the last three segments of the parameters in more detail and detected a delay of one minute between varying running speed and measured heart rate. Hence, the circulatory system of our population needed one minute to adapt to a change in running speed. The method we provided can be used to further investigate more HRV parameters.

KEYWORDS: ADAPTION OF HRV PARAMETERS, RUNNING, UNIVARIATE ANOVA WITH REPEATED-MEASURES, MULTIVARIATE ANOVA WITH REPEATED MEASURES

Introduction

The normal oscillation of the heart rate is called Heart Rate Variability (HRV). HRV is a measure which describes the parasympathetic and sympathetic influence of the autonomic nervous system. The interest in HRV has increased in the last few decades (Kaikkonen, Nummela, & Rusko, 2007). It is known that its influence changes under different conditions such as rest, physical exercise, mental stress, and body posture changes (supine, sitting, standing). Results how its influence changes during exercise have been inconsistent (Boettger, Puta, Yeragani, Donath, Müller, Gabriel, & Bär, 2010; Tulppo, Mäkikallio, Takala, Seppänen, & Huikuri, 1996).

The calculation of HRV parameters can be done in the time- or frequency-domain. In the time-domain, the heart rate at any point in time or the intervals between successive QRS complexes are determined (Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996). The intervals between two consecutive R peaks are called RR intervals. This equals the time between two heartbeats. Recommended time-domain parameters are the standard deviation of the RR intervals (SDRR), the integral of the density distribution (i.e. the number of all RR intervals) divided by the maximum of the density distribution (the HRV triangular index), the standard deviation of the average RR interval calculated over short periods (SDARR), and the square root of the mean squared differences of successive RR intervals (RMSSD) (Task Force, 1996).

In the frequency-domain, methods like the power spectral density (PSD) are used to obtain information of how power distributes as a function of frequency. Recommended frequency-domain parameters are the power within the very low frequency band $\leq$ 0.04 Hz), the power within the low frequency (LF) band (0.04 Hz to 0.15 Hz), the power within the high frequency (HF) band (0.15 Hz to 0.4 Hz), and the LF/HF ratio (Task Force, 1996).

The purpose of this study was to examine if and how three HRV parameters of the time- and frequency-domain (average heart rate, RMSSD, and LF/HF ratio) and one kinematic parameter (average running speed) are changing during one hour of running. We used datasets of 295 athletes where each dataset had a total length of about 65 minutes. Data was divided in segments of five minutes and the parameters were calculated for each segment. We applied two different analysis of variance (ANOVA) models to analyse the differences in the means of each segment for every parameter. The two ANOVA models were univariate ANOVA with repeated measures and multivariate ANOVA with repeated measures. The obligatory post-hoc procedure consisted of multiple dependent t tests with Bonferroni correction.

Methods

Hardware

We used the Polar RS800 Running Computer with an S3 stride sensor and a chest strap (Polar Electro Oy ("Polar"), Kempele, Finland). With this system, the running speed, the stride frequency, the barometric height, the heart rate, and the RR intervals were measured. The resolution of the RR intervals was 1 ms. The sampling frequency of the other four parameters was set to 0.2 Hz.

The complete measurement equipment consisted of three different sensor systems (Eskofier, Hoenig, & Kuehner, 2008, Eskofier, Kugler, Melzer, & Kuehner, 2012). The first sensor system was the Running Computer with the stride sensor and the chest strap. The second sensor system was an adidas_1 running shoe (adidas AG, Herzogenaurach, Germany). The third sensor system was the Nokia 6110 Navigator cell phone (Nokia, Espoo, Finland). In the current study, only data from the first sensor system (the Running Computer) was used.

Data

The study consisted of 431 runners whose running experience varied within the subjects. In the current study, we used a subset of only 295 subjects (98 female and 176 male[*], age 43 ± 11 years[*], BMI: 23.1 ± 2.4 kg/m^2[*], mean ± SD (standard deviation)). For the current study, only the running speed, the heart rate, and the RR intervals were used. In the subset of 295 subjects, all three variables were available.

The subjects got the task to complete a run in one hour in a self-determined fashion, without distance or speed requirements. Hence, the subjects could choose their own speed and

[*]27 subjects did not answer the questionnaire with respect to gender, age, and BMI.

therefore the intensity of the running exercise varied amongst subjects. The subjects ran for 60 minutes and data were acquired for 65 minutes. The subjects were free to run or to rest in the last 5 minutes of data acquisition (after the one-hour run), just as they preferred.

Analysis

Several parameters were calculated to obtain different HRV measures. All physiological parameters for HRV analysis were calculated from the tachogram (Figure 1), in which consecutive RR intervals are plotted (Task Force, 1996). The RR intervals were divided in segments of five minutes. In the time-domain, the average heart rate and the square root of the mean squared differences of successive RR intervals (RMSSD) were calculated (Task Force, 1996):

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{N-1} (RR_{i+1} - RR_i)^2}{N-1}} \quad , \tag{1}$$

where N is the number of RR intervals in one segment, RR_i (RR_{i+1}) is the i-th ((i+1)-th) RR interval in the specific segment and i ranges from 1 to N.

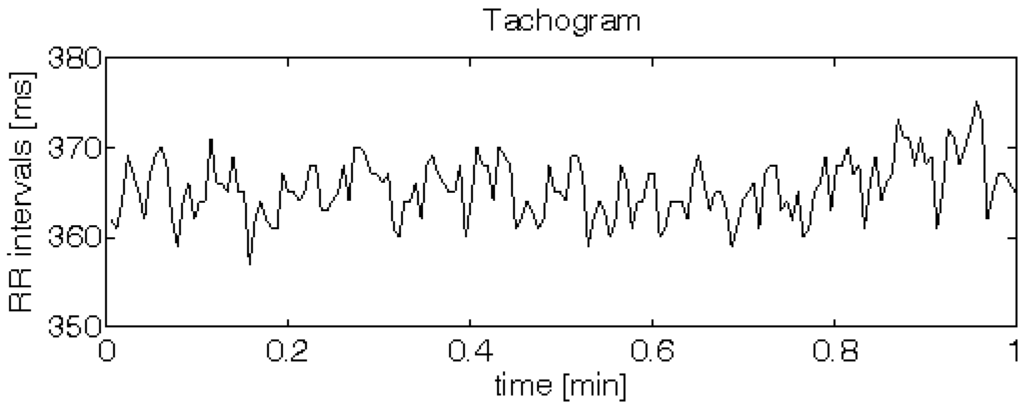

Figure 1. Exemplary tachogram of RR intervals in ms over time.

The RMSSD measure is known to be an adequate parameter for the analysis of measurements under uncontrolled conditions such as ‘free-running’ conditions (Penttilä, Helminen, Jartti, Kuusela, Huikuiri, Tulppo, Coffeng, & Scheinin, 2001, Plews, Laursen, Kilding, & Buchheit, 2012).

In the frequency-domain, the PSD of RR intervals was determined. After eliminating the DC component by subtracting the mean of the five minute segment, the Fourier transform (FT) of the signal was obtained. Then, the squared magnitude values of the FT were calculated. This resulted in the PSD. The calculation of the FT required equidistant values of RR intervals. Therefore the RR intervals were linearly interpolated with a sampling frequency of 8 Hz (Singh, Vinod, & Saxena, 2004). The PSD was normalized with the total power minus the power of the very low frequencies $\leq$ 0.04 Hz) (Task Force, 1996). The LF component of this PSD reached from 0.04 Hz to 0.15 Hz, and the HF component from 0.15 Hz to 0.40 Hz. The results of the parameter in the frequency-domain are based on the LF/HF ratio.

The dataset of the running speed was divided in segments of five minutes each likewise the tachogram. The average running speed within each segment was calculated as fourth parameter.

Statistics

All parameters were evaluated using a univariate analysis of variance (ANOVA) with repeated measures and a multivariate ANOVA with repeated measures (Stevens, 1996). In an ANOVA model, it is tested if the null hypothesis is accepted at the significance level α. Our null hypothesis was the assumption that there is no difference in the means (for each parameter over all subjects) of all segments. The alternative hypothesis is that at least two means differ significantly. If the null hypothesis was rejected, multiple dependent t tests with Bonferroni correction as post-hoc procedure were used. The Bonferroni correction was necessary to keep the overall α under control.

The basic requirements in applying one of these ANOVA models were independence of observations and multivariate normality. The independency was given due to the problem. When parameters after visualization in a histogram did not resemble a normal distribution, the natural logarithm of these parameters over the data set was used. The Lilliefors test (Lilliefors, 1967), which is a specialized version of the Kolmogorow-Smirnow test, was used at the significance level α for testing normal distribution. The method of univariate ANOVA with repeated measures assumes the sphericity or circularity assumption. If the sphericity assumption is violated, the Greenhouse & Geisser correction (Stevens, 1996) was applied to decrease the degrees of freedom.

Each analysis was performed using the Matlab package (MathWorks Inc., USA).

Results

The results section has been divided into two parts. The first part deals with the experiments described in the methods above. Due to these results, a further experiment with only the last three segments was performed. The results of this experiment are described in the second part.

Table 1. Results of the Lilliefors test and both ANOVA models. Abbr.: seg. = segment; uni. = univariate; multi. = multivariate; *The numbers in the brackets indicate segments in which the Lilliefors test was rejected.

	Lilliefors test (α = 0.10)	**uni. ANOVA (α = 0.05)**	**multi. ANOVA (α = 0.05)**
Heart rate	accepted (2)*	$F(1,294) = 1334.8$, $p < 0.001$	$F(12,283) = 480.2$, $p < 0.001$
ln(RMSSD)	rejected	n.a.	n.a.
ln(LF/HF)	accepted (1,13)*	$F(1,294) = 294.7$, $p < 0.001$	$F(12,283) = 126.5$, $p < 0.001$
Speed	accepted (1,9,11)*	$F(1,294) = 2194.3$, $p < 0.001$	$F(12,283) = 338.3$, $p < 0.001$

Table 1 states the results of the experiments over the complete data set. The three physiological parameters heart rate, ln(RMSSD), and ln(LF/HF) and the kinematic parameter speed were tested for the requirements of both ANOVA models. Although the Lilliefors test was not satisfied for every segment, both ANOVA models with repeated measures were applied to the parameters (McDonald, 2009) except for ln(RMSSD). Here no single segment fulfilled the requirement of multivariate normality. The sphericity assumption, necessary for the univariate

ANOVA model, was rejected for the three remaining parameters. Hence, the degrees of freedom for the univariate ANOVA were decreased with the Greenhouse & Geisser correction in these cases. In our case, both degrees of freedom were divided by the factor 12. Both ANOVA models revealed the same result. The null hypothesis of equal means within all 13 segments was rejected for every parameter.

Hence, the post-hoc procedure of multiple dependent t tests with Bonferroni correction was applied. The post-hoc procedure, presented in Figure 2, revealed that each parameter had adjacent segments with no significant differences.

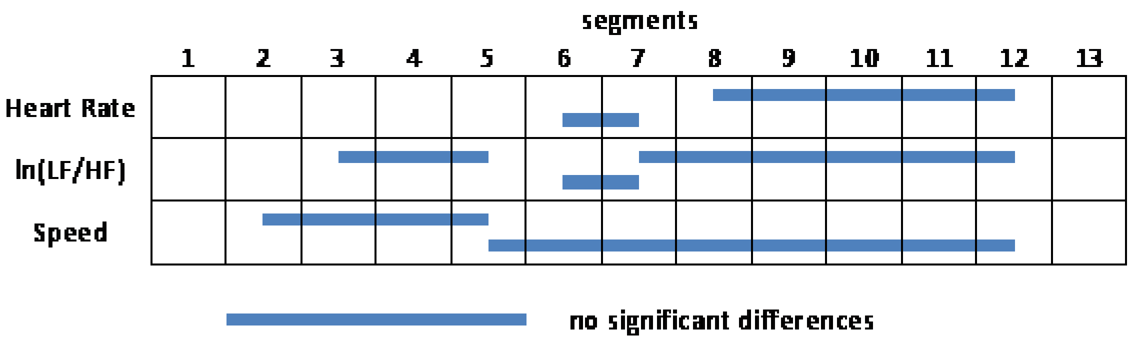

Figure 2. Results of the post-hoc procedure (multiple dependent t test with Bonferroni correction) for all three parameters. The blue bar indicates adjacent segments with no significant differences.

Because of the results, further investigations were done with the last three segments (segments 11 to 13). As each segment consisted of 5 minutes, the segments were divided in minutes. This resulted in a considered time of 15 minutes in total. Table 2 shows the results of the Lilliefors test and both ANOVA models. The parameter ln(RMSSD) was again tested for the requirements of the ANOVA models and, as neither the multivariate normality nor the sphericity assumption was fulfilled, not considered in further analysis. The three remaining parameters accepted mostly the normality assumption. The sphericity assumption was rejected for the three parameters, wherefore the Greenhouse & Geisser correction was used in the univariate case, too. Both ANOVA models revealed the result that the means of each parameter considering these 15 minutes were not equal. The post-hoc procedure is shown in Figure 3.

Table 2. Results of the Lilliefors test and both ANOVA models for the last three segments (last 15 minutes). Abbr.: seg. = segment; uni. = univariate; multi. = multivariate; *The numbers in the brackets indicate segments in which the Lilliefors test was rejected.

	Lilliefors test ($\alpha = 0.10$)	**uni. ANOVA ($\alpha = 0.05$)**	**multi. ANOVA ($\alpha = 0.05$)**
Heart rate	accepted (5)*	$F(1,294) = 213.4$, $p < 0.001$	$F(14,281) = 75.6$, $p < 0.001$
ln(RMSSD)	rejected	n.a.	n.a.
ln(LF/HF)	accepted (4,6,12,13)*	$F(1,294) = 70.1$, $p < 0.001$	$F(14,281) = 26.7$, $p < 0.001$
Speed	accepted (1,2,8,12)*	$F(1,294) = 450.7$, $p < 0.001$	$F(14,281) = 65.4$, $p < 0.001$

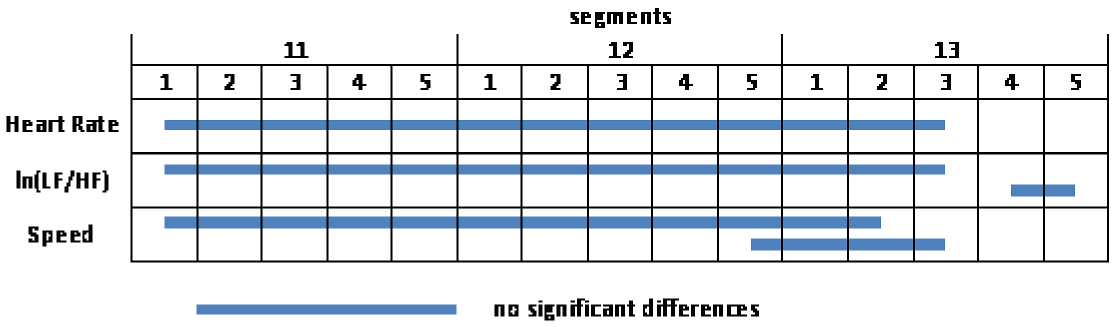

Figure 3. Results of the post-hoc procedure (multiple dependent t test with Bonferroni correction) for the last three segments (last 15 minutes) for three parameters. The blue bar indicates adjacent segments resp. minutes with no significant differences.

Discussion

The post-hoc procedure of multiple dependent t tests demonstrated that no different means for the segments 8 to 12 were present in all four parameters. As each of the three parameters changed from the 12th to the 13th segment, these segments were further investigated. The 13th segment is the last segment in which data was recorded and all volunteers were free to run or to rest, just as they preferred.

Partitioning the last three segments into five sub-segments of one minute length allowed a closer investigation of the end of the training session. Considering the three longest bars in Figure 3, the speed changed between the second and the third minute of the 13th segment. The heart rate and the ln(LF/HF) changed one minute later: between the third and the forth minute of the last segment. Here a delay of one minute was obvious.

The cardiovascular system has to deal with an increased demand during physical exercise (Aubert, Seps, & Beckers, 2003). As soon as the physical activity is finished, the cardiovascular system adapts to the current physiological demand. We found a delay of one minute between the physiological parameters and the kinematic parameter after decreasing the running speed. One minute is the time that our running population needed to adapt the circulatory system to a change in running speed.

Looking at the start of the training, the speed did not change between the 2nd to the 5th segment and then between the 5th to the 12th segment. The parameter ln(LF/HF) did not change between the 3rd to the 5th segment and between the 6th and the 7th segment. If a connection between these two parameters was existent, is uncertain.

The heart rate did not change between the 6th and the 7th segment. As the heart rate changed from the start of the exercise until the 6th segment, this is another evidence that the circulatory system needs time to adapt to physiological changes.

One drawback of this study is that neither a resting phase at the beginning nor the end had been included. Therefore, it is not possible to compare the physical exercise part with rest data or with recovery data. Another disadvantage is that the datasets were divided in segments of five minutes. Whether this length of the segments reflects well on the autonomous nervous system has to be further evaluated. Moreover, we only examined three HRV parameters and one kinematic parameter. One of the HRV parameters (RMSSD resp. ln(RMSSD)) had to be excluded as the requirements for the ANOVA models were not fulfilled. In further analysis, it is suggested to compare more recommended HRV parameters of the time- and frequency-domain (Task Force, 1996).

Conclusion and Outlook

We presented an analysis of the variation of heart rate and the ln(LF/HF) as well as the kinematic parameter running speed during a free one hour outdoor run. The ln(RMSSD) as a parameter for the chosen ANOVA models had to be excluded as the requirements for the ANOVA models were not fulfilled. Our analysis was based on two different ANOVA models with repeated measures. The used post-hoc procedure consisted of multiple dependent t tests with Bonferroni correction.

During this one hour of running, all three parameters reached a process in which the means did not alter significantly. We detected a delay of one minute between varying running speed and measured heart rate.

In further analysis, subgroups of athletes like female and male runners, or experienced and unexperienced runners will be examined. With these further investigations more information regarding HRV and fatigue could be gained.

Acknowledgments

We thank the adidas AG for financial and technical support of the study. This work was further supported by the Bavarian Ministry for Economic Affairs, Infrastructure, Transport and Technology and the European fund for regional development.

References

Lees, A., Graham-Smith, P. & Fowler, N. (1994). A biomechanical analysis of the last stride, touchdown, and takeoff characteristics of the mens long jump. *Journal of Applied Biomechanics*, *10* (1), 61–78.

Aubert, A.E., Seps, B. & Beckers, F. (2003). Heart Rate Variability in Athletes. *Sports Medicine, 33* (12), 889–919.

Boettger, S., Puta, C., Yeragani, V.K., Donath, L., Müller, H.-J., Gabriel, H.H.W. & Bär, K.-J. (2010). Heart Rate Variability, QT Variability, and Electrodermal Activity during Exercise. *Medicine and Science in Sports and Exercise*, *42* (3), 443–448.

Eskofier, B., Hoenig, F. & Kuehner, P. (2008). Classification of Perceived Running Fatigue in Digital Sports. *19th International Conference on Pattern Recognition* (pp. 1–4), Tampa,USA.

Eskofier, B., Kugler, P., Melzer, D. & Kuehner, P. (2012). Embedded Classification of the Perceived Fatigue State of Runners. *Proceedings of the IEEE International Conference on Body Sensor Networks* (pp. 113–117), London, UK.

Hottenrott, K., Hoos, O. & Esperer, H.D. (2006). Heart Rate Variability and Physical Exercise. Current Status. *Herz*, *31* (4), 544–554.

Kaikkonen, P., Nummela, A. & Rusko, H. (2007). Heart rate variability dynamics during early recovery after different endurance exercises. *European Journal of Applied Physiology*, *102* (1), 79–86.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, *62,* 399–402.

McDonald, J.H. (2009). *Handbook of Biological Statistics* (2nd ed.). Baltimore, Maryland: Sparky House Publishing.

Penttilä, J., Helminen, A., Jartti, T., Kuusela, T., Huikuiri, H.V., Tulppo, M.P., Coffeng, R. & Scheinin, H. (2001). Time domain, geometrical and frequency domain analysis of cardiac vagal outflow: effects of various respiratory patterns. *Clinical Physiology*, *21* (3), 365–376.

Plews, D.J., Laursen, P.B., Kilding, A.E. & Buchheit, M. (2012). Heart rate variability in elite triathletes, is variation in variability the key to effective training? A case comparison. *European Journal of Applied Physiology, 112* (11), 3729–3741.

Singh, D., Vinod, K., & Saxena, S. C. (2004). Sampling frequency of the RR interval time series for spectral analysis for heart rate variability. *Journal of Medical Engineering & Technology, 28* (6), 263–272.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Erlbaum.

Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). Heart Rate Variability – Standards of measurements, physiological interpretation, and clinical use. *European Heart Journal, 17* (3), 354–381.

Tulppo, M.P., Mäkikallio, T.H., Takala, T.E.S., Seppänen, T. & Huikuri, H.V. (1996). Quantitative beat-to-beat analysis of heart rate dynamics during exercise. *American Journal of Physiology, 271*, 244–252.

Music and Sound in (Exer)Games

Josef Wiemeyer

Institute of Sport Science, Technische Universität Darmstadt, Germany

Abstract

Music plays in important role in human culture. This is also the case for exercise, games and exergames. In this study we examine the influence of previous game experience and the presence of music and sound on game performance and game experience. Sixteen students (12 males, 4 females; age: M = 24.8 years, SD = 3.4) played two exergames with and without sound and music. Game performance, game experience and the perception of sound and music were assessed as dependent variables. Whereas music and sound had no impact on game performance, selected dimensions of game experience (tension, positive affects) were significantly influenced. The game-specific significance of music and sound was confirmed by the participants.

KEYWORDS: MUSIC, SOUND, EXERGAMES, GAME EXPERIENCE, SERIOUS GAMES

Introduction

Music – as an "integral – and often inescapable phenomenon of … culture" (Bishop, 2010, p.35) – is an important part of human lives. In Germany, for example, about 90 per cent of young people (age: 12 to 19 years) are interested in music. After "love & friendship", "music" is the second important interest of young people (MPFS, 2011). Not only children and youth, but also adults spend a considerable amount of time to listening to music (Statistisches Bundesamt, 2004).

Music has important influences on human cognitive, emotional and perceptual-motor functions (for a critical review, see Hunter & Schellenberg, 2010):

> Music captures attention, raises spirits, triggers a range of emotions, alters or regulates mood, evokes memories, increases work output, heightens arousal, induces states of higher functioning, reduces inhibitions and encourages rhythmic movement (Karagheorgis & Priest, 2012a, p.35).

Table 1 illustrates the various general impacts of sound and music on different levels of the human system.

Music has been applied in sport and exercise to enhance performance, for example when learning and performing perceptual-motor skills (e.g., Beisman, 1967; Bishop, 2010) or exercising (Karageorghis & Priest, 2012a and b). For instance, stimulating music may increase strength and endurance performance.

Table 1. Impact of music on different organizational levels of humans (synopsis from literature; references: see text).

Level	Impact
Cognitive	Attention (focus, distraction), perception, memory
Emotional	Mood, emotion, motivation, adhesion, volition
Physiological	Arousal (stimulation), relaxation (sedation), disinhibition; heart rate, skin conductance, EMG
Behavioral	Work output, rhythmic movement (timing), precision, movement speed

Music can be applied asynchronously, i.e. pre-task or post-task, or synchronously, i.e., in-task. The following sport-specific effects are ascribed to the asynchronous use of music:

- Transitional benefits: arousal control, reduced rating of perceived exertion (RPE), improved mood, ergogenic effects, reduced muscle tension
- Chronic benefits: Increased exercise adherence, effective preparation routine

The synchronous application of music refers to “an innate human predisposition to synchronize movement with musical rhythms” (Karageorghis & Priest, 2010, p.49). Although this explanation is mainly a phenomenal description, research in neuroscience has revealed some interesting phenomena of afferent-efferent synchronization. These effects can be used to enhance timing and precision of movements.

Generally, the impact of music on humans is subject to numerous moderators including characteristics of the music (e.g., beat, melody, tempo, volume) and the listener (e.g., personality, mood, music preference, performance level) as well as the type of task to be completed (e.g., strength, endurance, precision, speed, intensity) and the mode of application (e.g., pre-task, in-task, schedule). In Figure 1 the complex interaction of various factors and moderators is illustrated.

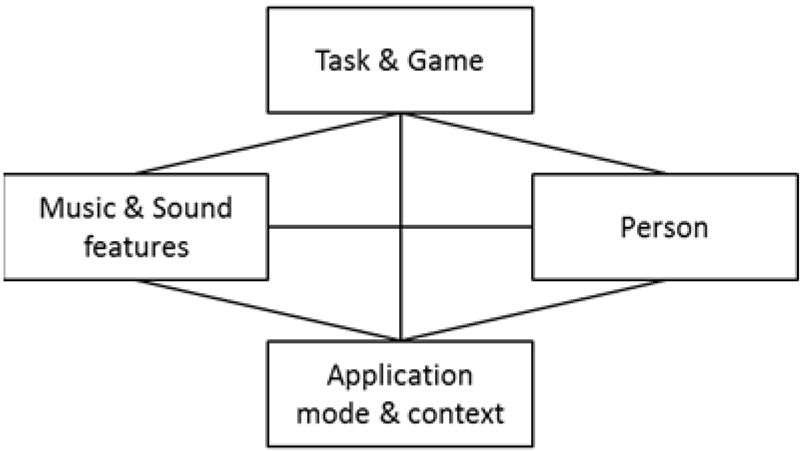

Figure 1. The impact of music as a complex interaction of various factors.

Music and sound are also important features in digital games. In two experiments, Nacke

(2009) confirmed significant effects of diegetic sound and non-diegetic music on Game Experience (GX) and physiological measures when playing first-person shooter (FPS) games. GX was measured by the Game Experience Questionnaire, whereas physiological reactions were assessed by electro-dermal activity (skin conductance) and the EMG of facial muscles (Mm. orbicularis oculi, corrugator supercilii, zygomaticus major). The presence of diegetic sound had the strongest impact on GX. Furthermore, besides general effects of music on GX and EMG, Nacke (2009) found specific gender effects: Female players experienced significantly less negative affects, greater challenge (only younger women) and less arousal, when music was present. Diegetic sound and nondiegetic music show a complex interaction: When diegetic sound is present in FPS games, music does not play a significant role.

No studies are known to the author analyzing specific effects of music and sound on performance in exergames, i.e., digital games including whole-body movements and exercises. Therefore, the purpose of this study was to confirm the generic (nondiegetic) and explore the specific (diegetic) effects of music and sound on game performance in exergames. This generalization is on the one hand supported by Mueller et al. (2011), who address the impact of external rhythm (music) on exergame performance. Evidence also comes from movement science, where the use of acoustic support like (rhythmic) sonification has been proved to enhance both perception and performance of movements (e.g., Effenberg, 2005).

Methods

Hypotheses

Based on existing evidence the following hypotheses were derived:

- Music and sound have a differential impact on game performance and experience depending on its guiding, i.e. diegetic versus nondiegetic function. Analogous to the results reported by Nacke (2009) and due to the guiding function of diegetic music and sound (e.g., Effenberg, 2005) we expect a stronger effect of diegetic music and sound compared to a weaker effect of nondiegetic music and sound both on game performance and game experience.
- Previous experience on digital games modifies music perception in exergames. Due to the novelty effect we expect music perception to be more pronounced in participants without previous game experience.

Sample

Based on the results of Nacke (2009) yielding effect sizes of music and sound versus no music/sound ranging from 0.22 (experiment 1: $N = 36$) to 0.60 (experiment 2: $N = 36$), a mean effect size of 0.41 was calculated. Optimal sample size was calculated using the program GPOWER 3.1 (Faul et al., 2007). Considering an α error of 0.05, an β error of 0.20 (power = 0.80) and an effect size of 0.41 an optimal sample size of 14 resulted for a within-subjects design with two repeated measures (correlation between the two measurements = 0.5).

Sixteen students (12 males, 4 females; age: M = 24.8 years, SD = 3.4) volunteered to participate in the study. 68.8% (8 males) are experienced videogamers (criteria: playing digital games for more than five years), 31.3% have played the Nintendo Wii and 18.8% have played the Balance Board.

Design and Dependent Measures

The study was based on a between-subjects design with repeated measures. Previous game experience (cutoff point: 5 years of game experience) and game played (Wii fit Step-up versus Hula Hoop) were the two between-subjects factors.

Dependent measures were game score, game experience, general and game-specific perception of music and sound.

Game experience was assessed using a modified 35-item version of the Game Experience Questionnaire (Poels et al., 2008; German translation: Nacke, 2009). To offer a uniform response format for all questionnaires, the items were estimated using a 7-point Likert scale (1 – "fully applies"; 7 – "does not apply at all"). With one exception (negative affects) reliability was acceptable, i.e., above .7 (see Table 2).

Table 2. Reliability of the 6 GEQ dimensions (Cronbach's alpha).

Condition	GEQ dimension					
	Flow	Competence	Tension	Challenge	Positive affect	Negative affect
Music	.791	.938	.865	.792	.787	.421
No music	.724	.935	.825	.772	.879	.162

The general significance and experience of music and sound was assessed based on the Brunel Music Rating Inventory-2 (BMRI-2; Karageorghis et al., 2006). We applied a modified 13-item questionnaire with a 7-point Likert scale (1 – "fully applies"; 7 – "does not apply at all"). Six items addressed the impact of music (BMRI-2) and seven items were dedicated to sound.

Game-specific perception of music and sound was assessed using a self-developed questionnaire comprising 4 items per game with a 7-point Likert scale (1 – "fully applies"; 7 – "does not apply at all"; see Appendix).

Procedure

The participants were randomly (i.e., by lot) assigned to a Step-up group (n=8; 5 males, 3 females) or a Hula Hoop group (n=8; 7 males, 1 female) with each group consisting of an equal number (n=4) of experienced and less experienced videogamers, respectively. After completing a questionnaire to collect basic data like age and previous game experience the participants played either the Wii game 'Step-up' or the Wii game 'Hula Hoop'. Loudness of sound and music was at a convenient level. No physical measurements were performed.

The two games were chosen as the most appropriate games from the Wii fit exergame collection because they demand dynamic whole-body activities and because music and sound have different functions. The 'Step-up' game (see Figure 2a) is a kind of step-aerobic where the players have to step on and off the balance board according to a prescribed rhythm. In this game music and sound exert strong and permanent diegetic functions. Music supports feedforward control of timing, whereas sound conveys feedback about correct timing. In addition, timing is indicated by visual cues presented synchronously to the music. In the 'Hula Hoop' game (see Figure 2b) the task is to rotate as many hoops as possible within a certain time. Music is nondiegetic, whereas sound indicates flight of the hoops thrown at irregular intervals by the two avatars in the background, thus exerting a weaker and transitory diegetic function compared to the 'Step-up' game. The hoops must be caught by the player by leaning to the side.

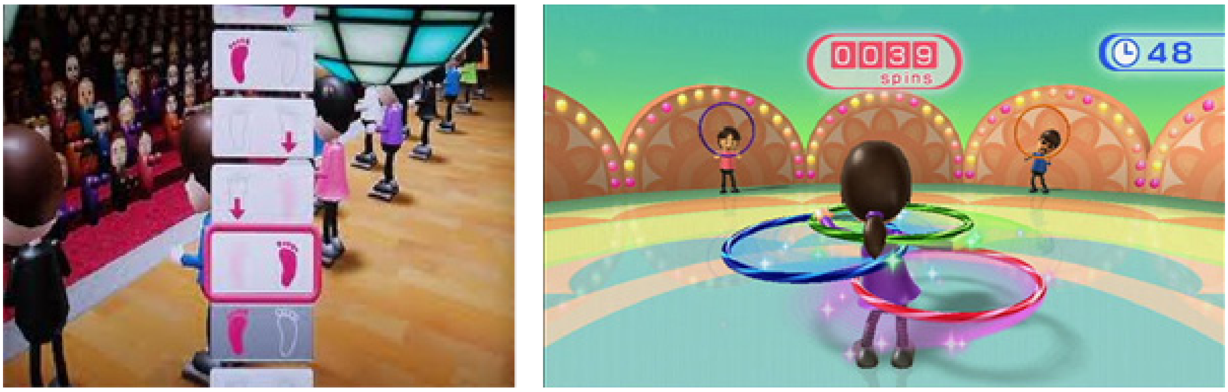

Figure 2. Illustration of the Wii games 'Step-up' (left side) and 'Hula Hoop' (right side).

Table 3 illustrates the significance of music and sound in the two games.

Table 3. Significance of music and sound in the two Wii games.

Game	Music	Sound
Step-up	Diegetic: timing, feedforward (strong; permanent)	Diegetic: timing, feedback (strong; permanent)
Hula Hoop	Nondiegetic	Diegetic: flight (weak; transitory)

The participants had to play the games once (Step-up) or twice (Hula Hoop) with and without music and sound. After playing, GX, game performance, and the significance of music and sound were assessed as dependent variables.

Data Analysis

Statistical analysis was performed using the SPSS 20.0 package. Two or three factor ANOVAs were used with repeated measures (music) and one or two between-subjects factors (game experience, game). Follow-up analyses were performed using either Wilcoxon tests or Mann-Whitney U tests.

Results

Concerning game score (Figure 3) a 2 (previous game experience) x 2 (game) x 2 (music) ANOVA with repeated measures on the last factor revealed a significant main effect of the game factor ($F_{1,12}$ = 7.79, p<0.05, η^2_{part} = 0.39). Furthermore, neither main effects nor interactions were significant.

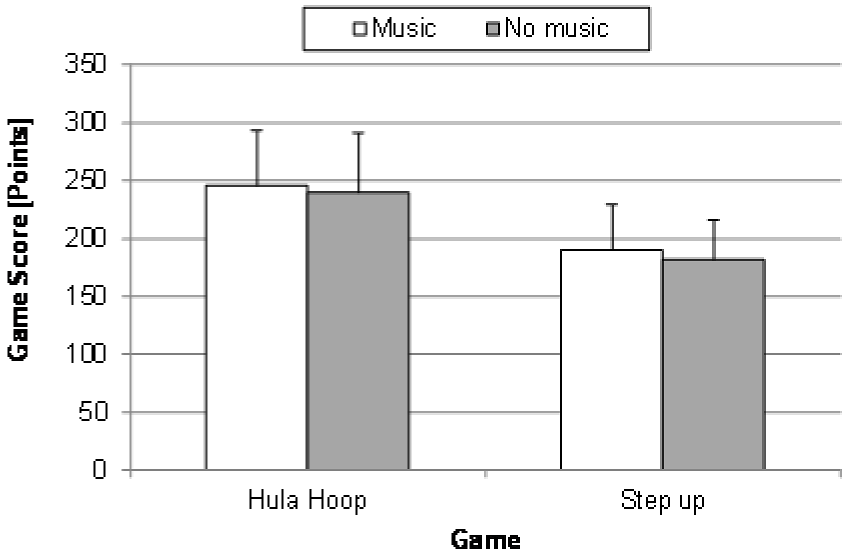

Figure 3. Game scores with and without music (Mean, Standard Deviation).

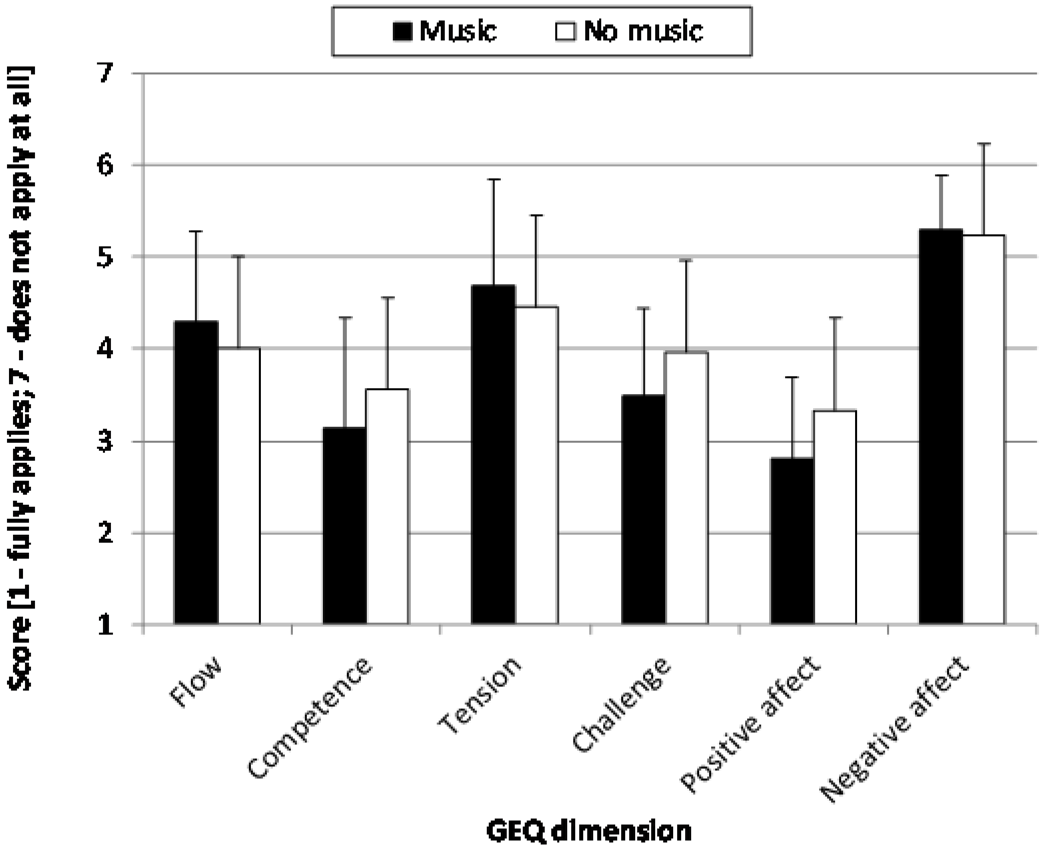

Figure 4. GEQ scores with and without music (Mean, Standard Deviation).

The GEQ items generally showed intermediate levels of GX ranging from 2.8 to 5.3 (music) and 3.3 to 5.2 (no music), respectively (Figure 4). A 2 (previous game experience) x 2 (game) x 2 (music) x 6 (GEQ dimensions) ANOVA with repeated measures on the last two factors revealed a significant main effect of GEQ dimension ($F_{2,24} = 16.15$; $p<.001$, $\eta^2_{part} = 0.57$; after correction with $\varepsilon_{Greenhouse\text{-}Geisser}$) and an interaction of music and GEQ dimension ($F_{3,32} = 3.45$; $p<.01$, $\eta^2_{part} = 0.22$; after correction with $\varepsilon_{Greenhouse\text{-}Geisser}$; Figure 4). Follow-up analyses (Wilcoxon tests) revealed significant differences both in favor of and against music and sound (music and sound: higher tension, higher positive affect).

All other main effects and interactions were not significant.

Concerning subjective experience of music and sound a 2 (previous game experience) x 2 (game) x 13 (items) ANOVA with repeated measures on the last factor revealed neither significant effects of game nor previous game experience. However, a significant effect of items was found ($F_{4,47} = 10.35$; $p<.001$, $\eta^2_{part} = 0.46$; after correction with $\varepsilon_{Greenhouse\text{-}Geisser}$; Figure 5). In general, the impact of sound was more agreed upon than the influence of music. Agreement for cognitive and motivational effects of sound was highest. The strongest performance effect of music was perceived for the beat, followed by rhythm, tempo, and melody.

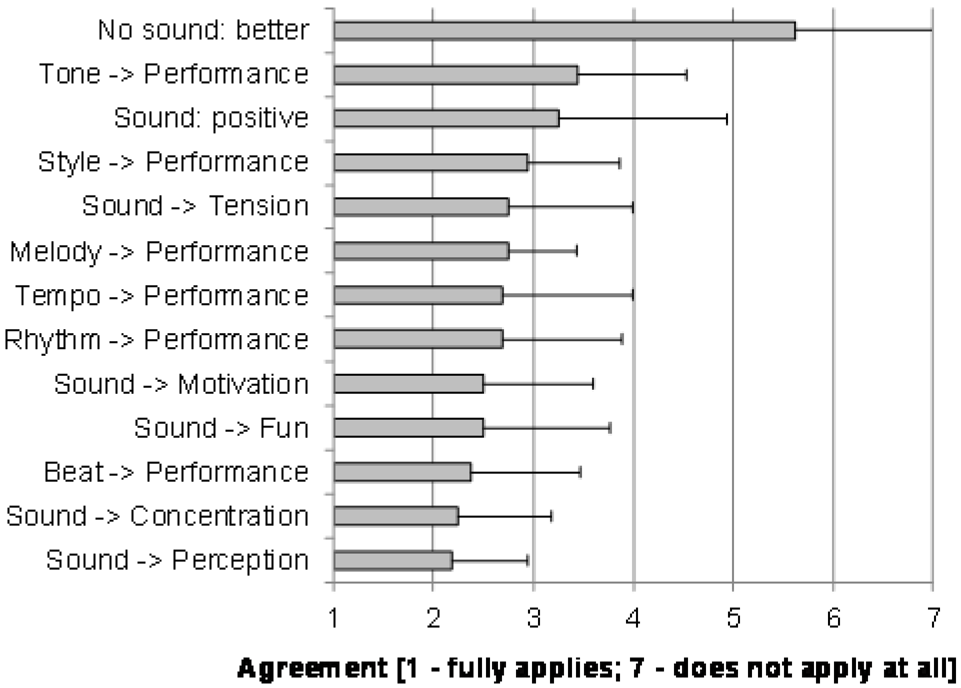

Figure 5. General significance of sound and music (Mean, Standard Deviation).

Two 2 (previous game experience) x 4 (items) ANOVAs did not reveal any main effects nor interactions for the game-specific effects of sound and music. In the 'Step-up' game two items displayed agreement (support of resumption and feedback of timing; Figure 6) whereas the other two items were only moderately in agreement (guidance function in general and in case of errors).

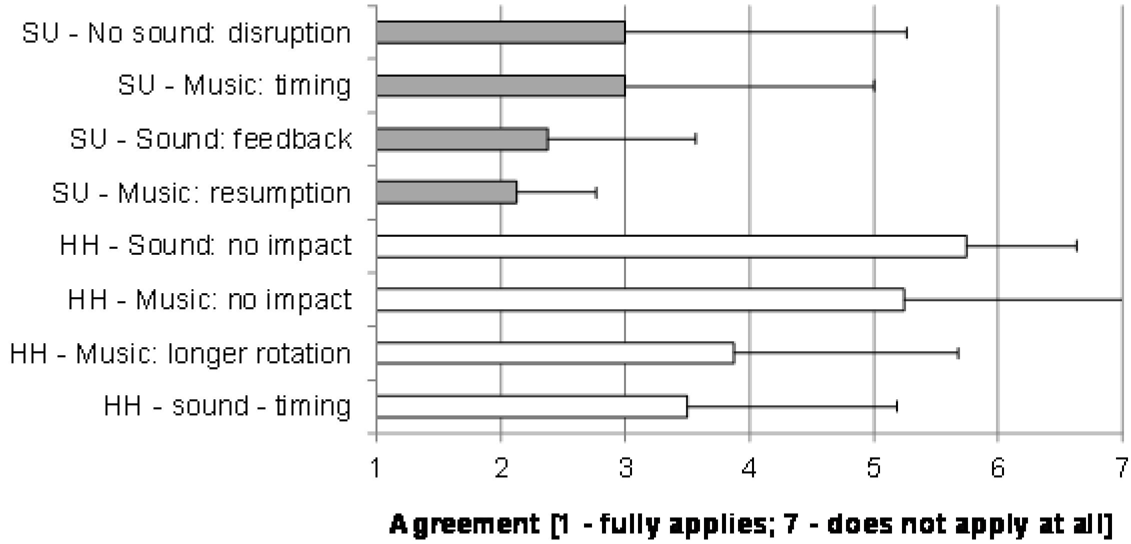

Figure 6. Game-specific significance of sound and music (Mean, Standard Deviation) – HH – Hula Hoop; SU – Step-up.

In the 'Hula Hoop' game a moderate general influence of music was experienced (2 items), whereas no particular ergogenic or guidance effect was reported.

Discussion

We were not able to show differential effects of music and sound in the two games on performance. The participants scored equally well with and without music and sound. The missing performance effects of music and sound may be due to the fact that the participants had only one or two trials with the respective game. Playing time is always a compromise of getting accustomed to and engaged in the game and losing interest. We expected that the first one or two attempts would be most important for the impact of music and sound.

Second, overall game score is a comparatively superficial indicator of performance. More sophisticated measures like Absolute and Variable Errors might have yielded more differentiated results. Another reason may be that the missing information conveyed by sound and music could be compensated or substituted by visual sources in both games. A different result may have been obtained when the participant would have played blindfolded in the sound/music condition, but this was not the research question of the present experiment. The present study was performed to investigate the additional effect of sound and music in exergames.

The significantly higher score found in the Hula Hoop game is due to the specific scoring systems of the games.

However, we found positive and negative effects of music and sound on GX. Regardless of the game and game experience music and sound enhanced tension and positive affects. This result corresponds to the findings reported by Nacke (2009). However, Nacke also found significant effects on competence, flow, negative affects, and challenge. Again, the lower impact of missing acoustics may be due to the fact that visual sources of information could be used by the players.

Actual power of this study – based on the GEQ items (mean effect size = -0.14; mean correlation = 0.65) – was poor (1-β = 0.24). Therefore, a total sample size of 74 participants would have been required to confirm the impact of music and sound.

Furthermore, despite the absence of behavioral effects, the participants confirmed specific support for continuous exercising (music) and feedback (sound) in the Step-up game and a moderate influence of sound on timing in the Hula Hoop game. Specific cognitive effects were more approved than global performance effects. Thus, there is only limited support for hypothesis 1.

Due to different scales the GEQ results of our study cannot directly be compared to the results of Nacke (2009). Nacke (2009) applied a five-point scale (0 – not at all; 4 – extremely) whereas our study included a seven-point scale (1 – fully applies; 7 – does not apply at all) in order to establish uniform scales in all questionnaires. Reliability of the GEQ items competence, tension, and challenge was higher in our study with lower reliability of flow, positive and negative affects. Overall the moderate effects of gaming on GX correspond to the results proposed by Nacke (2009), who found low to moderate levels of GX when playing a first person shooter video game on the Wii console. However, there are also particular differences: Nacke (2009) found higher levels of flow, competence, and challenge. Further studies must show whether the genre of exergames generally elicits lower degrees of GX as compared to other game genres. Problems with long-term motivation in exergames may be considered another indicator that moderate GX may be a specific problem of exergames (Wiemeyer, 2010).

One shortcoming of this study was the fact that the influence of music and sound could not be separated. Therefore, only the combined influence of ‘sound and music’ (both diegetic in the ‘Step-up’ game; sound diegetic in the ‘Hula Hoop’ game) could be studied.

Hypothesis 2 could not be confirmed. Experience with games seems not to have a general impact on new gaming activities. Either the cutoff point (5 years of game experience) was not appropriate or quality rather than quantity of game experience plays the more important role. It is reasonable but remains an open question whether domain-specific transfer is possible as has been found in balance training (Kliem & Wiemeyer, 2010).

Further research should use more game trials, a separation of music and sound, more sophisticated performance parameters, and more detailed measures of game experience.

Conclusion

The study confirms significant effects of sound and music on selective aspects of game experience in exergames. With music and sound tension and positive affects are higher and specific diegetic functions are perceived more pronounced than global functions. Although performance effects could not be found, these results suggest that developers of (exer)games should deliberately choose music and sound that has diegetic functions supporting players in a more specific way compared to nondiegetic function which seems to be less significant when diegetic music and sound are present. For players of (exer)games the results imply that by turning music and sound on and off they can change the difficulty of the game and add another source of variable gaming.

Acknowledgement

The author would like to thank Lars Engelhardt and Jakub Pilarski for their valuable support concerning planning and execution of the study.

References

Beisman, G.L. (1967). Effect of rhythmic accompaniment upon learning of fundamental motor skills. *Research Quarterly*, *38* (2), 172–176.

Bishop, D. (2010). 'Boom Boom How': Optimising performance with music. *Sport and Exercise Psychology Review*, *6*, 35–47.

Effenberg, A.O. (2005). Movement sonification: Effects on perception and action. *IEEE Multimedia. Special Issue on Interactive Sonification, 12* (2), 53–59.

Faul, F., Erdfelder, E., Lang, A. & Buchner, A.-G. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39* (2), 175–191.

Hunter, P.G. & Schellenberg, E.G. (2010). Music and emotion. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception* (pp. 129–164). New York: Springer.

Karageorghis, C. & Priest, D.-L. (2008). Music in Sport and Exercise: An Update on Research and Application. *The Sport Journal*, *11* (3).

Karageorghis, C.I. & Priest, D.-L. (2012a). Music in the exercise domain: a review and synthesis (Part I). *International Review of Sport and Exercise Psychology*, *5* (1), 44–66.

Karageorghis, C.I. & Priest, D.-L. (2012b). Music in the exercise domain: a review and synthesis (Part II). *International Review of Sport and Exercise Psychology*, *5* (1), 67–84.

Karageorghis, C.I., Priest, D.-L., Terry, P.C., Chatzisarantis, N.L.D. & Lane, A.M. (2006). Redesign and initial validation of an instrument to assess the motivational qualities of music in exercise: The Brunel Music Rating Inventory-2. *Journal of Sports Sciences*, *24* (8), 899–909.

Kliem, A. & Wiemeyer, J. (2010). Comparison of a Traditional and a Video Game-Based Balance Training Program. *International Journal of Computer Science in Sport*, *9* (Special Edition 3: Serious Games), 78–89.

MPFS (Eds.). (2011). JIM-Studie 2011. Jugend, Information, (Multi-)Media. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger in Deutschland. Stuttgart: MPFS. [Online-Version]

Mueller, F., Edge, D., Vetere, F., Gibbs, M.R., Agamanolis, S., Bongers, B. & Sheridan, J.G. (2011). Designing Sports: A framework for exertion games. *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada.

Nacke, L. E. (2009). *Affective Ludology: Scientific Measurement of User Experience in Interactive Entertainment*. Blekinge Institute of Technology, Doctoral Dissertation Series No. 2009:04.

Poels, K., de Kort, Y.A.W. & IJsselsteijn, W.A. (2008). FUGA - The fun of gaming: Measuring the human experience of media enjoyment. Deliverable D3.3: Game Experience Questionnaire. TU Eindhoven, Eindhoven, The Netherlands.

Statistisches Bundesamt (Eds.). (2004). Alltag in Deutschland. Analysen zur Zeitverwendung [Everyday in Germany. Analysis for time use]. Wiesbaden: Statistisches Bundesamt.

Wiemeyer, J. (2010). Gesundheit auf dem Spiel? – Serious Games in Prävention und Rehabilitation [Health in risk? – Serious games in prevention and rehabilitation]. *Deutsche Zeitschrift für Sportmedizin*, *61* (11), 252–257.

Appendix: Self-developed questionnaire (German version)

Questions concerning “Hula Hoop”

	trifft voll zu	trifft zu	trifft etwas zu	unentschlossen	trifft eher nicht zu	trifft nicht zu	trifft gar nicht zu
	1	2	3	4	5	6	7
1. Der Sound beim Spiel "Hula Hoop" hatte auf mich keinen Einfluss.							
2. Die Musik hilft mir beim Spiel "Hula Hoop" die Reifen länger zu drehen.							
3. Beim "Hula Hoop" ohne Sound wusste ich nicht wann ich die Reifen fangen sollte.							
4. Die Musik beim "Hula Hoop" hatte auf mich keinen Einfluss.							

Questions concerning "Step-up"

	trifft voll zu	trifft zu	trifft etwas zu	unentschlossen	trifft eher nicht zu	trifft nicht zu	trifft gar nicht zu
	1	2	3	4	5	6	7
1. Beim "Step Up" mit Musik weiß ich genauer wann ich auf das Board auftreten soll.							
2. Beim "Step Up" ohne Sound komm ich schnell aus der vorgegebenen Schrittreihe raus.							
3. Wenn ich beim "Step Up" einen Fehler mache, komme ich mit Musik wieder schneller rein.							
4. Der Sound beim Spiel "Step Up" gibt mir ein Feedback wie gut ich auf dem Balanceboard aufgetreten bin.							