Journal of Sports Science 9 (2021) 9-16
doi: 10.17265/2332-7839/2021.01.002

Reviving Direct Observation Methods for Physical Activity Behavior

Pedro Miguel Ribeiro da Silva[1], Sérgio Hélder da Silva Soares Soares[2], Jorge Augusto Pinto Silva Mota[1], Paula Maria Marques Moura Gomes Viana[2, 3], and Pedro Miguel Machado Soares Carvalho[2, 3]

1. CIAFEL, Faculty of Sports, University of Porto, Porto, 4200-450, Portugal

2. School of Engineering, Polytechnic of Porto, Porto, 4200-072, Portugal

3. INESC TEC, Porto, 4200-465, Portugal

Abstract: Purpose: Current Direct Observation (DO) methods are burdensome to the observer and relevant limitations can be identified on the outcome information. This study aims to characterize DO from the observers' point of view and to analyze the feasibility and usefulness of visual information in the form of video. Method: data collection was done with ten adult males, observed while performing different physical activities in an indoor court. Sessions were video recorded by two cameras. Three observers performed systematic observation, using SOPLAY, with different sampling rates. Inter observer's agreement and with reference data was analyzed by Cohen's Kappa statistic. Results: SOPLAY highest agreement between observers was 0.611 on vigorous category and walking category had the lowest 0.188. It doubles the time needed to annotate the video with pauses, but half of player's behavior is preserved. Conclusion(s): using video to support DO permits to collect more accurate data and a more detailed behavior categorization is warranted. Developments in computer vision are expected to reduce the human efforts in DO methods.

Key words: Measuring physical behavior, systematic observation, movement intensity, video

1. Introduction

Public health surveillance will undoubtedly benefit from having accurate data concerning physical activity (PA) and sedentary behavior (SB) in the population. A better understanding of the factors that influence active lifestyles will help to improve the effectiveness of any public health activity programming [1]. The levels of PA performed, at any point of an individual lifespan, reflect a complex interaction of biological, psychological and sociological factors [2]. Therefore, assessing PA and SB is a challenging task and many different approaches have been proposed.

Advances in technologies and modeling techniques have also led to the development of new pattern recognition approaches and devices that provide enhanced ways for monitoring and evaluating PA and SB [3]. Currently, the combination of sensors (such as accelerometers and heart rate monitors) has become standard and is capable of measuring total physical activity as well as components of physical activity that play important roles in human health [4]. Nevertheless, direct observation (DO) methods are still considered as the most effective (gold standard) technique because behavior is directly observed [5]. One of the advantages of DO techniques is capturing the context of the observed behavior, which instruments, such as accelerometer and heart rate monitors, can not provide [6].

However, diverse sources of error can be identified when using DO: (1) if individual behavior is of interest, one observer is needed to code it and sometimes more than one observer is needed to classify just one participant behavior (e.g. SOFIT)—these methods require considerable manpower; (2) researcher's burden is also high, because current methods require tedious manual

Corresponding author: Pedro Miguel Ribeiro da Silva, PhD, research fields: Physical Activity, Sedentary behavior, Performance, Methods of monitoring behavior, Health and Active technology.

coding, which impose a great barrier when used for large scale monitoring. This means that usually only a small number of participants can be observed; (3) the most common method of DO uses time sampling technics and the moment of observation (usually 10 seconds) is used to reflect the behavior until the next observation period. This results in a low sampling frequency; (4) to reduce the subjectivity associated with human observers, few behavior categories can be considered (e.g. 3 in SOPLAY—Sedentary, Walking and Vigorous); (5) to try to keep observer's objectivity, the time of the observation session can not be long.

Even though current DO techniques enable coding both the behavior and the context in which it occurred, these methods are underused by researchers [5]. This is mainly due to the significant limitations they impose, such as the cost and burden associated with collecting and processing data. This barrier typically leads to a limited scope or amount of observation data, bringing into question whether data from those samples are only adequate for that setting or whether they can be generalizable. Hence, there is a need to explore the improvement of DO tools that may facilitate the collection of space-time information about PA and SB in a more cost-effective manner and at higher sampling frequencies than current DO methods.

The present study aims to: (1) characterize DO from the observers' point of view and to analyze the feasibility and usefulness of visual information in the form of video; (2) demonstrate relevant sources of error that may occur by comparing observer's classifications; (3) enhance the role that video surveillance can play to assist researchers when using DO to assess PA and SB.

2. Methods

2.1 Participants

Participants in this study were a convenience sample since the video recording system was installed in a private gym. Ten adult males were simultaneously observed while performing different physical activities. Each participant freely performed distinct activities during the entire session; no restrictions were imposed on their behavior. The study protocol was approved by the Portuguese Foundation for Science and Technology and the CIAFEL research center; it was also obtained individual informed consents from the participants in order to conduct the study. Even though we only used adults in this study, due to privacy and legal constraints, the analysis performed is also valid and applicable for children.

2.2 Video Sequences Acquisition

A set of video sequences capturing scenes depicting different participants, performing various movements and activities, were acquired for the purpose of analyzing the use of video in the characterization of PA. The observed individuals occupied two distinct areas of the space (half indoor court with 20 × 20 meters) each covered by a camera. In the experiments, two different IP cameras were used: Sony SNC-CH120 (Sony Corp., Japan) and Vivotek FD8162 (Vivotek, New Taipei City, Taiwan). The cameras were placed in a high position looking down, augmenting the area covered and minimizing occlusion situations. Even though the cameras were different, the videos were captured with similar characteristics: resolution of 1,024 × 768; 30 frames per second; MP4 codec. Fig. 1 depicts two frames, an image for each camera, for illustration purposes.

2.3 Direct Observation Methodology

The System for Observing Play and Leisure Activity (SOPLAY) was the instrument chosen to characterize participant's behavior, since it was designed to assess group levels of PA in different settings and environmental contexts [7], and due to its relevance in the context of PA characterization. It uses momentary group time sampling to code PA in three categories: sedentary, walking and vigorous, thus indicating the

Fig. 1 A frame perspective of the two cameras field of view.

percentage of participants attending or involved in different activity categories [8]. The analysis of the video sequences was performed by three observers well trained in the usage of the SOPLAY method. The process was conducted in 3 phases, to analyze different aspects of the method and usage scenarios and to examine the benefits of using video, for the purpose of activity observation and characterization. The adopted phases were:

(a) Annotation without pauses—this setup corresponds to the original SOPLAY method, where an observation period (10 seconds) is followed by a period of annotation (10 seconds). In this case, there is a loss of information corresponding to the time in which the observer annotates the visualization results of the previous time slot. Each video segment had the duration of 8:04 minutes and the total time to perform this annotation was 16:08 minutes.

(b) Annotation with pauses—the difference with respect to the original method consists in stopping the video during the period in which the annotation is made, avoiding the loss of information. The annotation time was 19:33 for the first video and 18:15 for the second, amounting to a total time of 37:48 (21:40 minutes more than the first setup).

(c) Reference annotation—observers had the possibility to pause and go back in the video to the start of the annotation to visualize again in order to resolve ambiguities and ensure consensus and accuracy of the observation. The annotation time in this setup was 37:12 for the first video and 27:03 for the second, amounting a total time of 64:15 seconds (48:07 minutes more than the first setup). This annotation was taken as reference for our analysis.

To avoid interferences between observers, observations were made individually and each phase was separated from the previous one by an interval of 2 weeks or more to avoid observer bias or fatigue. In the last stage, after an individual first observation, there was a discussion period between observers to ensure full compliance, resulting in the generation of the activity classification to be taken as the reference information. The form provided by SOPLAY manual was used for annotation purpose. One column was added to the original form to include the degree of confidence (using a scale from 1—the lowest confidence to 10—the highest confidence) of observers with regards to his classification.

3. Statistical Analysis

Observers' agreement was analyzed using different approaches. First, the agreement between each observer individually and the reference annotation was assessed. Secondly, the agreement within observers was evaluated. The following measures were computed: Precision—number of classifications

compliant with the reference; Precision with agreement—number of classifications compliant with the reference and with at least another observer; Error rate—number of classifications not compliant with the reference and without the agreement of other observers; Error rate with agreement—number of classifications not compliant with the reference, but with the agreement of at least another observer.

To conclude, the agreement between observers using the Cohen's Kappa statistic was calculated. The strength of agreement (Kappa value) was interpreted as follows (Altman, 1991): < 0.20 poor; 0.21-0.40 fair; 0.41-0.60 moderate; 0.61-0.80 good; 0.81-1.00 very good.

All analyses were performed by using the software Microsoft Excel 2010 (Microsoft Office for Windows, version 2010; USA) and the Statistical Package for Social Sciences (IBM Statistical Package for Windows, version 21.0; USA); the level of significance was set at $p \leq 0.05$.

4. Results

Results of the video sequences analysis performed by the three observers, in each of the 3 types of annotations, are presented in the following tables: Table 1, assessment of annotations without pauses; Table 2, assessment of annotations with pauses; Table 3, assessment of annotations with and without pauses over the same intervals, i.e., considering only the intervals of observation without pauses. In the first annotations, where there is no video pause, the original SOPLAY method resulted in 52 annotations to characterize behavior for each of the three observers. In the second annotation, with pauses, it doubled the amount of annotations, 104 observations for each observer.

Tables 1-3 depict curious results, especially a precision lower than expected. The highest precision value found was 57.60% (Table 1, Observer 2). The precision decreases when agreement between observers is also required—the highest value was 44.23% (Table 1, Observer 2). Results from Tables 1 and 2 show an improvement in accuracy and a decrease of the amount of information lost. Even though the average precision decreases when performing annotation with pauses, this is due to higher error for observer 2; the two other observers depict an increase in precision.

In Tables 4 and 5 we can see that observers tend to have greater confusion when people classify the behavior "walking", leading to incorrect characterization of the motion. In the event of vigorous movement, the greatest confusion is with the movement in which the person is walking and also in counting the number of people observed.

With the purpose of understanding the level of certainty that observers had in each behavior characterization was also assessed the degree of confidence (in a scale from 1 to 10) that each observer considered for every annotation.

Table 1 Assessment of the SOPLAY video annotation without pauses.

Percentage (%)	Observer 1	Observer 2	Observer 3	Mean
Precision	50.00	57.69	36.54	48.08
Precision with agreement	38.46	44.23	34.62	39.10
Error rate	40.38	25.00	51.92	39.10
Error rate with agreement	9.62	17.31	11.54	12.82

Table 2 Assessment of the SOPLAY video annotation with pauses.

Percentage (%)	Observer 1	Observer 2	Observer 3	Mean
Precision	55.77	40.38	42.31	46.15
Precision with agreement	41.35	36.54	33.65	37.18
Error rate	27.88	41.35	46.15	38.46
Error rate with agreement	16.35	18.27	11.54	15.38

Table 3 Comparison of the SOPLAY video annotation with and without pauses.

	Observer 1		Observer 2		Observer 3		Mean	
Percentage (%)	no pause	with pause	no pause	with pause	no pause	with pause	no pause	with pause
Precision	50.00	53.85	57.69	44.23	36.54	46.15	48.08	48.08
Precision with agreement	38.46	42.31	44.23	38.46	34.62	38.46	39.10	39.74
Error rate	40.38	32.69	25.00	36.54	51.92	40.38	39.10	36.54
Error rate with agreement	9.62	13.46	17.31	19.23	11.54	13.46	12.82	15.38

Table 4 Confusion matrix (percentage) of the observer's annotations without pauses (52 scans), on the three categories of SOPLAY (sedentary, walking and vigorous).

Annotations (%) confusion matrix	Sedentary			Walking			Vigorous		
	Obs. 1	Obs. 2	Obs. 3	Obs. 1	Obs. 2	Obs. 3	Obs. 1	Obs. 2	Obs. 3
As sedentary	82.7	80.8	69.3	5.8	5.8	7.7	0	1.9	0
As walking	7.7	9.6	19.2	82.7	82.7	76.9	1.9	0	5.8
As vigorous	7.7	7.7	3.8	7.7	7.7	3.8	90.4	90.4	80.7
Wrong number of participants	1.9	1.9	7.7	1.9	1.9	9.7	7.7	7.7	13.5
As the other two categories	0	0	0	1.9	1.9	1.9	0	0	0
Confidence scale (1-10)	7.9	5.9	5	7.9	5.9	5	8	6	5

Table 5 Confusion matrix (percentage) of the observer's annotations with pauses (104 scans), on the three categories of SOPLAY (sedentary, walking and vigorous).

Annotations (%) confusion matrix	Sedentary			Walking			Vigorous		
	Obs. 1	Obs. 2	Obs. 3	Obs. 1	Obs. 2	Obs. 3	Obs. 1	Obs. 2	Obs. 3
As sedentary	91.4	68.3	70.2	17.3	8.7	2.9	0	0	0
As walking	3.8	15.4	21.2	70.2	76.9	81.7	4.8	4.8	10.6
As vigorous	1	6.7	2.9	6.7	7.7	6.7	86.5	86.5	85.5
Wrong number of participants	3.8	5.8	1.9	5.8	6.7	7.7	8.7	8.7	2.9
As the other two categories	0	3.8	3.8	0	0	1	0	0	1
Confidence scale (1-10)	8	6.8	4.5	8	6.8	4.5	8.1	6.8	4.5

Table 6 Agreement values (Cohen's Kappa) between the three observers, on the three categories of SOPLAY annotation (sedentary, walking and vigorous). Top values are from the first annotation without pauses and values bellow the diagonal in bold are from the annotation with pauses.

Agreement (Cohen's Kappa)	Sedentary			Walking			Vigorous		
	Obs. 1	Obs. 2	Obs. 3	Obs. 1	Obs. 2	Obs. 3	Obs. 1	Obs. 2	Obs. 3
Observer 1		0.581	0.412		0.441	0.188		0.611	0.433
Observer 2	0.514		0.568	0.462		0.415	0.609		0.482
Observer 3	0.465	0.511		0.258	0.288		0.566	0.442	

Table 6 presents the highest agreement between observers was 0.611 in the vigorous category and this is considered moderate agreement. The lowest was in the walking category 0.188, considered poor. All agreement values were significant at $p < 0.001$. It is noteworthy that the higher values are on the extreme categories, sedentary and vigorous, and the lowest in the walking category.

5. Discussion

While direct observation methods are often considered as the gold standard to categorize PA behavior, they need to be delimited to their specific methodology. This means that the results emerging

from its use should be interpreted on the light of the chosen observation procedures and characteristics. The results of this study highlight some of the limitations of current direct observation methodology, in particular from the observers' point of view.

The selection of instruments for measuring PA depends on the PA component of interest, study objectives, characteristics of the target population and study feasibility in terms of cost and logistics [4]. An important decision to process the output data of any instrument used to assess PA, concerns to the number of codes that should be used to classify PA intensity, being cut-points in accelerometers or behavior characterization in direct observation. In direct observation method, the more codes one instrument has, the harder it will be to the observer to code PA, probably promoting higher levels of fatigue, which could be the cause of the decrease of the reliability of the measures, as our study demonstrated. Even if more codes would be associated with an increase in precision of the measure, the precision gained may not be necessary, depending on the research question being answered [7].

An important limitation of direct observation is the need to train observers because direct observation techniques strongly rely on the accuracy and skills of the observer to identify and classify PA behaviors. Inter-observer reliability has been tested and considered acceptable for both SOPLAY contextual variables and activity counts (IOA = 80%, R = 0.75) [9]. Observers must be properly trained to be objective and nonjudgmental, and steps must be taken to ensure they maintain their skills over time; there is also the possibility of observed people behaving differently when an observer is present (i.e. reactivity). In this study, observers were well trained following the guidelines of the SOPLAY authors to try to guarantee an adequate data collection. Nevertheless, our results presented unsatisfactory levels of precision and agreement between observers.

In direct observation, observational periods are usually divided into short observe and record moments; intervals are equally distributed between both time periods. The sampling method used will determine what participants have to watch, when to watch and how to record their behavior. Different instruments use different sampling techniques such as: momentary time sampling (instantaneous or scan sampling); partial time sampling (recording the event if it occurs at any time during the observe interval); and whole interval sampling (the event is recorded only if it occurs through the whole interval). Sampling periods are usually well defined, either using stop watches or audiotape players with pre-recorded signals to initiate and end recording periods [7].

The SOPLAY method used in this study is designed to capture random snapshots of activity levels and these snapshots are presumed to reflect activity behavior of the group during the whole period of time. However, identifying the frequency of the scans to accurately capture activity levels is still an unresolved question [10]. The reference annotation used in this study was held using the traditional SOPLAY methodology by three observers; the time sampling to categorize behavior was of 10 seconds. In these time windows, the categories have to be mutually exclusive. Consequently it is assumed that the behavior (sedentary, walking or vigorous) was sustained for 10 seconds. This assumption can be problematic in the case of younger children since they are prone to short bursts of activity [11, 12].

By performing direct observation over the recorded videos with different observers we were able to: (1) generate reference or ideal information, (2) assess the reliability of the classifications obtained, and (3) present the usefulness and feasibility of using video to increase observations accuracy. The results show that even well trained observers struggle to classify PA and SB, not only when compared to the ground truth but also they disagree frequently. Moreover, their confidence levels in performing the observations were

very different between the three observers, with the third observer showing an average confidence level of 5 in a scale of 10 points. Greater confusion was when they had to code "walking" behavior, leading to incorrect characterization of the motion. There was also confusion between sedentary and walking situations that might be explained due to the tenuous difference between them. This indicates the existence of a high degree of subjectivity associated with direct observation procedures (real time scanning ending in an instantaneous classification) prone to errors since it is not possible to revisit the observation time period. Moreover, half the behavior information is lost because of the annotation time period.

When using video, it is possible to avoid the need to have an observer present thus diminishing the intrusiveness, in the sense that there is no wearable device, but also because the participant's reactivity issue does not occur. It also avoids the loss of information and, even without going back in the video, an improvement in the precision of the results was observed. The augmented time required is directly related to the analysis of the complete video and not just half of it. The decrease of the pressure for a fast annotation is also reflected in a higher confidence. Hence, accurate results can be obtained not just by pausing the video, but also by rewinding it to eliminate any doubts.

Progresses in sensor, communications and computer technologies favor their introduction in the PA and SB assessment problem [3], as Pratt et al., 2012 [13] stated "the greatest potential to increase population physical activity might thus be in creation of synergistic policies in sectors outside health including communication and transportation". Therefore, future development of sensors and analytical techniques for assessing PA and SB should focus on the dynamic ranges of sensors, comparability for sensor output across manufacturers, and the application of advanced modeling techniques to predict energy expenditure and classify physical activities. Moreover, new approaches for qualitatively classifying PA should be validated using direct observation or recording [4].

New advances in automated video-based processing techniques offer considerable promise to overcome limitations of direct observation. For example, it may minimize or eliminate the observation load imposed on researchers by direct observation and it has the advantage of not being intrusive to the participants. In the case of the automatic observation (using tracking algorithms) these changes of behavior, even if in less than 10 seconds, are not ignored since the participant is tracked all the time at the established frame rate (for example 30 frames per second).

6. Conclusion

Tracking systems can be trained to follow individuals continuously, and depending on the frame rate capabilities of the video camera used to collect the data, we can have sampling rates from 30 frames per second to 120 or more frames per second with a relatively low budget video camera. Computer vision can offer new solutions to automatically characterize human behavior, such as PA and SB, and we consider it a logical step to advance this research field.

What Does This Article Add?

The results of this study highlight some of the limitations of current direct observation methodology, in particular from the observers' point of view. New advances in automated video-based processing techniques offer considerable promise to overcome limitations of direct observation. Computer vision can offer new solutions for automatically characterize human behavior, such as PA and SB.

Acknowledgements

The authors would like to acknowledge the contribution of Ana Carvalhinho and Simone Medeiros de Oliveira during the data collection of this study.

Funding Source

This work was funded by the Portuguese Foundation for Science and Technology (FCT), grant SFRH/BPD/71332/2010, Pest-OE/SAU/UI0617/2011 and UIDB/50014/2020. The work was also developed in the context of Project QREN 33910 ARENA, a R&D project funded by ERDF through ON2 as part of the NSRF, and managed by IAPMEI—Agência para a Competitividade e Inovação, I.P..

References

[1] Sallis, J. F., Prochaska, J. J., and Taylor, W. C. 2000. "A Review of Correlates of Physical Activity of Children and Adolescents." *Med Sci Sports Exerc* 32 (5): 963-75.

[2] Sallis, J. F., and Hovell, M. F. 1990. "Determinants of Exercise Behavior." *Exerc Sport Sci Rev* 18: 307-30.

[3] Silva, P., Andrade, M. T., Carvalho, P., and Mota, J. 2013. "A Structured and Flexible Language for Physical Activity Assessment and Characterization." *Journal of Sports Medicine* 9. doi: 10.1155/2013/420916.

[4] Butte, N. F., Ekelund, U., and Westerterp, K. R. 2012. "Assessing Physical Activity Using Wearable Monitors: Measures of Physical Activity." *Med Sci Sports Exerc* 44 (Suppl. 1): S5-12. doi: 10.1249/MSS.0b013e3182399c0e.

[5] McKenzie. 2010. "2009 C. H. McCloy Lecture. Seeing Is Believing: Observing Physical Activity and Its Contexts." *Res Q Exerc Sport* 81 (2): 113-22.

[6] McKenzie, Crespo, N. C., Baquero, B., and Elder, J. P. 2010. "Leisure-Time Physical Activity in Elementary Schools: Analysis of Contextual Conditions." *J Sch Health* 80 (10): 470-7. doi: 10.1111/j.1746-1561.2010.00530.x

[7] McKenzie. 2002. "Use of Direct Observation to Assess Physical Activity." In *Physical Activity Assessments for Health-Related Research*, edited by Welk, G. 179-95. Champaign, IL: Human Kinetics.

[8] McKenzie, Catellier, D. J., Conway, T., Lytle, L. A., Grieser, M., Webber, L. A., et al. 2006. "Girls' Activity Levels and Lesson Contexts in Middle School PE: TAAG Baseline." *Med Sci Sports Exerc* 38 (7): 1229-35. doi: 10.1249/01.mss.0000227307.34149.f3.

[9] McKenzie, Marshall, S. J., Sallis, J. F., and Conway, T. L. 2000. "Leisure-Time Physical Activity in School Environments: An Observational Study Using SOPLAY." *Prev Med* 30 (1): 70-7. doi: 10.1006/pmed.1999.0591.

[10] Saint-Maurice, P. F., Welk, G., Ihmels, M. A., and Krapfl, J. R. 2011. "Validation of the SOPLAY Direct Observation Tool with an Accelerometry-Based Physical Activity Monitor." *J Phys Act Health* 8 (8): 1108-16.

[11] Rowlands, A. V., Eston, R. G., and Ingledew, D. K. 1997. "Measurement of Physical Activity in Children with Particular Reference to the Use of Heart Rate and Pedometry." *Sports Med* 24 (4): 258-72.

[12] Welk, G. J., Corbin, C. B., and Dale, D. 2000. "Measurement Issues in the Assessment of Physical Activity in Children." *Res Q Exerc Sport* 71 (Suppl. 2): S59-73.

[13] Pratt, M., Sarmiento, O. L., Montes, F., Ogilvie, D., Marcus, B. H., Perez, L. G., and Brownson, R. C. 2012. "The Implications of Megatrends in Information and Communication Technology and Transportation for Changes in Global Physical Activity." *Lancet* 380 (9838): 282-93. doi: 10.1016/S0140-6736(12)60736-3.