Automated High-Frequency Observations of Physical Activity Using Computer Vision

JORDAN A. CARLSON[1,2], BO LIU[3], JAMES F. SALLIS[4], J. AARON HIPP[5], VINCENT S. STAGGS[2,6], JACQUELINE KERR[4], AMY PAPA[1], KELSEY DEAN[1], and NUNO M. VASCONCELOS[3]

[1]Center for Children's Healthy Lifestyles and Nutrition, Children's Mercy Hospital, Kansas City, MO; [2]School of Medicine, University of Missouri-Kansas City, Kansas City, MO; [3]Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA; [4]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA; [5]Department of Parks, Recreation, and Tourism Management and Center for Geospatial Analytics, North Carolina State University, Raleigh, NC; and [6]Biostatistics and Epidemiology, Health Services & Outcomes Research, Children's Mercy Hospital; Kansas City, MO

ABSTRACT

CARLSON, J. A., B. LIU, J. F. SALLIS, J. A. HIPP, V. S. STAGGS, J. KERR, A. PAPA, K. DEAN, and N. M. VASCONCELOS. Automated High-Frequency Observations of Physical Activity Using Computer Vision. *Med. Sci. Sports Exerc.*, Vol. 52, No. 9, pp. 2029–2036, 2020. **Purpose:** To test the validity of the Ecological Video Identification of Physical Activity (EVIP) computer vision algorithms for automated video-based ecological assessment of physical activity in settings such as parks and schoolyards. **Methods:** Twenty-seven hours of video were collected from stationary overhead video cameras across 22 visits in nine sites capturing organized activities. Each person in the setting wore an accelerometer, and each second was classified as moderate-to-vigorous physical activity or sedentary/light activity. Data with 57,987 s were used to train and test computer vision algorithms for estimating the total number of people in the video and number of people active (in moderate-to-vigorous physical activity) each second. In the testing data set (38,658 s), video-based System for Observing Play and Recreation in Communities (SOPARC) observations were conducted every 5 min (130 observations). Concordance correlation coefficients (CCC) and mean absolute errors (MAE) assessed agreement between (1) EVIP and ground truth (people counts+accelerometry) and (2) SOPARC observation and ground truth. Site and scene-level correlates of error were investigated. **Results:** Agreement between EVIP and ground truth was high for number of people in the scene (CCC = 0.88; MAE = 2.70) and moderate for number of people active (CCC = 0.55; MAE = 2.57). The EVIP error was uncorrelated with camera placement, presence of obstructions or shadows, and setting type. For both number in scene and number active, EVIP outperformed SOPARC observations in estimating ground truth values (CCC were larger by 0.11–0.12 and MAE smaller by 41%–48%). **Conclusions:** Computer vision algorithms are promising for automated assessment of setting-based physical activity. Such tools would require less manpower than human observation, produce more and potentially more accurate data, and allow for ongoing monitoring and feedback to inform interventions. **Key Words:** DIRECT OBSERVATION, BUILT ENVIRONMENT, PARK, SCHOOL, VIDEO

Physical activity occurs in multiple settings, including parks and schoolyards. The environmental features within these settings can support or inhibit physical activity (1). Thus, public health researchers and practitioners seeking to improve understanding of environmental influences on physical activity, evaluate physical activity interventions, and track population physical activity trends in such settings commonly use ecological (group level) physical activity assessment tools such as the System for Observing Play and Recreation in Communities (SOPARC) (2–7). These tools use momentary direct observation (4), which involves visiting a community setting (e.g., park) and conducting an environmental scan of users in the area to document the activity level of each user, at the moment they are observed, as sedentary, walking, or vigorous (8).

There are several limitations to existing tools. The scans can be challenging when people's activity changes rapidly and when a large number of people are in the setting and crossing paths with one another. In-person observation requires substantial staffing, training, observation time, and data entry/management, which limits the scalability of these tools. Both inadequate training of staff conducting the scans and inadequate frequency of data capture can result in invalid data (9,10). The momentary snapshots of behavior often do not generalize to a person's activity during their entire time in the setting or to other people's activity in the setting (11). Perhaps most importantly, assessing only a snapshot (moment) of physical activity substantially limits these tools' utility for intervention monitoring and feedback. Moving to automated high-frequency data capture would enable continuous assessment and allow researchers, program leaders, and decision makers to make rapid data-informed decisions to support just-in-time interventions (12,13).

Address for correspondence: Jordan A. Carlson, Ph.D., 610 E. 22nd St. Kansas City, MO 64108; E-mail: jacarlson@cmh.edu.
Submitted for publication December 2019.
Accepted for publication March 2020.

0195-9131/20/5209-2029/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000002341

Computer vision, a rapidly growing field of artificial intelligence that uses deep learning models to train computer algorithms to classify features within images and video (14), provides the opportunity to automate ecological physical activity assessment. Although research has shown that computer vision is promising for assessing physical activity (15), valid computer vision–based ecological physical activity assessment tools do not yet exist. Such tools are needed because they could provide more accurate and representative (over time) estimates of physical activity due to high-frequency data capture, and likely result in greater use of ecological physical activity assessment (improved scalability).

The purpose of this study was to develop and test the validity of the Ecological Video Identification of Physical Activity (EVIP) computer vision algorithms for analyzing video to estimate group-level physical activity in park and school settings. Ecological Video Identification of Physical Activity provided similar outputs as provided by SOPARC, including the number of people in the target area and number of people physically active.

METHODS

Participants and Procedures

Data were collected from nine sites during organized activities, including after-school programs, park-based programs, and Physical Education classes. Across the nine sites, the following types of settings were captured: four open green spaces/sports fields, two paved surfaces, two gymnasiums, and two playgrounds (examples in Figure 1). Data collection occurred across 1–4 visits per site (22 total visits), with each visit lasting approximately 1 h. All people involved in the organized activities were enrolled in the study and outfitted with an accelerometer. Two cameras captured video recordings of the target area in which the activity occurred. The number of participants during each visit ranged from 9 to 56, and most were youth. A mean of 179.1 min (SD = 87.3 min) of data was collected at each site, for a total of 1611.9 min (26.9 h). This study was approved by the sponsoring institution's human subjects' protection committee and participants provided informed consent.

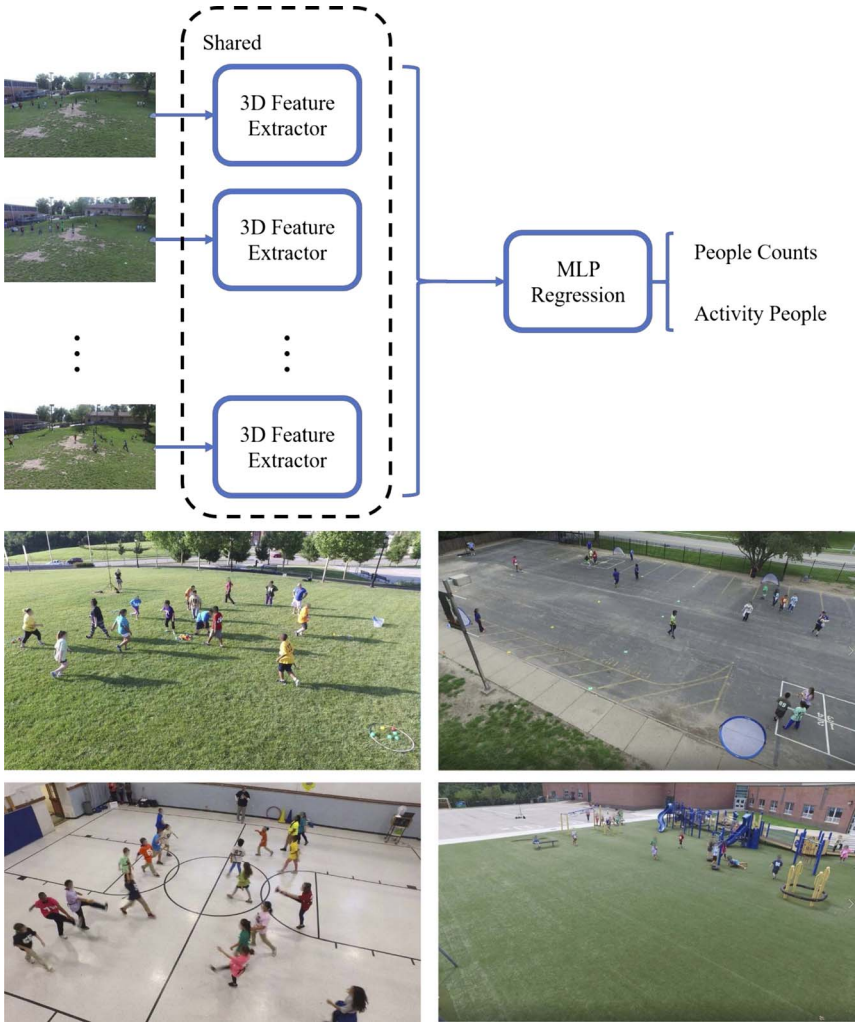

FIGURE 1—Computer vision system for EVIP and example video screen shots. Note: Image from left to right, top to bottom are open green space/sports field, paved surface, gymnasium, and playground.

Measures

Video. Two DJI Osmo 4K wide angle cameras (Model X3/FC350H; SZ DJI Technology Co.; Ltd, Shenzhen, China) were used to capture the video recordings. Each camera was affixed to a Studio Assets MegaMast camera tripod that was able to be telescoped to a height of 27.5 ft (Koll Ltd; Chicago, IL). An Apple iPad (Apple Inc., Cupertino, CA) was used to control the camera via the DJI GO application. Each camera captured a different perspective of the target area and was moved halfway through each visit to capture a second perspective. Thus, a total of four videos capturing unique perspectives were collected during each visit, with the exception of three visits that were not long enough to capture the additional two perspectives. Across the 22 visits, 82 videos capturing unique perspectives were collected. The camera height and distance from the base of the camera to the nearest/bottom part of the target area captured in the video ranged across videos from 8.5 to 24.4 ft (median = 14.9 ft) and 5.3 to 35.4 ft (median = 14.9 ft), respectively. After data collection, research staff viewed each of the 82 videos and rated the percent of the scene that was obstructed (e.g., due to trees or buildings) and the percent that was covered by shadows (i.e., appearing darker due to the sun being shadowed by trees or buildings).

Counting number of people (ground truth measure for number of total people). Because the settings captured involved organized activities, the same number of people were generally in the target area for a given study visit. The target areas were restricted during data collection to minimize the occurrence of nonparticipants entering the area, and participants were instructed to stay within the area captured by the camera. However, in some circumstances, it was difficult to prevent participants from entering and exiting the viewing area. Thus, research staff viewed each video after the visits to identify all instances of when a person left and/or entered the viewing area and record this information in a database. To facilitate this process, participants wore T-shirts of various colors with distinct numbers on the front and back. This allowed the research staff to denote that a given participant's accelerometer data (i.e., linked to their shirt number and color) should be set as "missing" for the second they left the scene until the second they returned. The number of people with nonmissing accelerometer data for a given second served as the ground truth for the number of total people in the scene.

Accelerometers (used to provide ground truth measure for number of active people). Participants wore an ActiGraph GT3X accelerometer (ActiGraph LLC; Pensacola, FL) on a belt at their left iliac crest, with vertical axis counts derived for 1-s epochs. The Freedson youth 3 METs (metabolic equivalents) age-based cutpoint was used to classify each second as moderate-to-vigorous physical activity (MVPA) or not MVPA (16). The cutpoint for 12-yr-olds was applied to all participants because participants' ages were not collected and 12 yr reflected the midpoint in the age range covered by the Freedson equation. We chose the Freedson 3-METs cutpoint because of its lower counts threshold than other cutpoints (17), which was preferred because it better matched the SOPARC walking category as walking can include light activity (8). Derived variables were created by aggregating the data across participants who were in the scene (based on the "counting number of people" methods described above) to derive the total number of active (in MVPA), each second. The purpose of using second-level activity level information was to provide high-frequency ground truth information for training the EVIP algorithms.

SOPARC observations. To identify whether EVIP had a similar level of validity as direct observations conducted by humans (i.e., current standard of practice for field-based observations), SOPARC observation scans were conducted approximately once every 5 min on the testing video data, for a total of 130 observations. The SOPARC scans captured the total number of people in the scene and the number active (walking and vigorous were combined) (8). Each scan took between 4 and 44 s to complete (mean = 20.7 s). The video was not stopped or rewound during the observation to mimic in-person observations. The observations were completed by two raters who were trained using the SOPARC training materials (8). A subset of 26 observations were completed independently by both raters, and inter-rater agreement for the number of people active was good (ICC = 0.86).

Training and Testing Data Sets

The videos were divided into 1-s clips, with 60% of the clips from each site being randomly allocated to a training data set and the remaining 40% to a testing data set. The training data set comprised 57,987 s from all 82 videos (camera perspectives) from all 22 visits representing all nine sites, and the testing data set comprised 38,658 s from 78 videos (camera perspectives) from all 22 visits representing all nine sites.

Computer Vision Algorithm Development

The EVIP algorithms were developed by adapting existing validated computer vision modules for action recognition. The system is illustrated in Figure 1. First, the C3D feature extraction module was used to extract local visual and motion features. It had been pretrained on a large-scale activity data set (18) and was transferred to the EVIP data by fine-tuning. It implements a fully convolutional 3D neural network to extract features from the video clips (19). Because the module uses deep learning, features were not determined *a priori* but rather are extracted using layers of multiplicative information and thus are not always well understood or interpretable. Next, the output from the feature extraction module was used in a multilayer perceptron regression, which was trained for multitask regression to predict both the number of total people in the scene and number of people active (in MVPA) using a loss function that optimized both predictions.

Analyses

For both the number of people in the scene and number of people active, boxplots were used to examine the distribution of EVIP error (predicted − ground truth) across the range of ground truth values in the second-level data set. Agreement between EVIP and ground truth measurements was assessed

by computing the mean absolute error (MAE) and the value of Lin's concordance correlation coefficient (CCC), which measures goodness of fit around the line of perfect agreement (20). To compute confidence limits for the MAE and CCC, a nonparametric bootstrap procedure was carried out. A bootstrap sample of videos was constructed by randomly selecting 78 videos with replacement, then randomly selecting one 1-s clip from each of the selected videos. The MAE and CCC were computed for each of 2000 such bootstrap samples, and the bootstrap MAE and CCC distributions were used to compute 95% confidence limits. This procedure was chosen to minimize the effect of clustering of clips within videos while capturing the variability in sampling both videos and clips within videos. The agreement analyses described above were repeated on an aggregated data set created by dividing the data into 30-s intervals and averaging the EVIP and ground truth measurements across each such interval. Where clips could not be divided evenly into 30-s intervals, the final interval was less than 30 s. Intervals covering less than 10 s were excluded from analysis.

To examine potential sources of bias, associations of site and scene features with EVIP error (predicted − ground truth) were estimated by fitting linear mixed models. Camera height, camera distance, percent of scene obstructed, and percent of scene covered by shadows were explanatory variables in one model, and setting type (open green space/sports field, paved surface, gymnasium, or playground) was the explanatory variable in the second model, fitted separately due to correlations between setting type and other explanatory variables. Each model included a random intercept for video number to adjust for clustering of clips within videos. The bootstrap procedure described above was used to obtain bootstrap samples, the mixed models for the two types of EVIP error (number in scene, number active) were fitted to each bootstrap sample, and the bootstrap distributions of the regression coefficients were used to obtain 95% confidence limits.

To compare the performance of EVIP and SOPARC observations in predicting the number of people in the scene and number active, EVIP and ground truth estimates were averaged across each SOPARC observation time period. Mean absolute errors and CCC were computed to assess agreement between each test method and ground truth (predicted − ground truth). Bootstrapping was used to construct 95% confidence limits for MAE and CCC by resampling from the 130 direct observations.

Criteria for interpreting and CCC were: small (≤0.40), moderate (0.41–0.60), large (0.61–0.80), and very large (0.81–1.0) (21). Statistical analyses were performed in R (22).

RESULTS

According to ground truth measurements for the 38,658 clips in the testing data, the median number of total people in a scene was 16 (interquartile range [IQR], 11–23; range, 0–50), and the median number of people active in a scene was 4 (IQR, 2–8; range, 0–32; Table 1). The CCC representing agreement between EVIP and ground truth for the number of people in the scene was very large (CCC = 0.88), and the MAE was 2.70 people (16.9% of the ground truth median). The CCC representing agreement between EVIP and ground truth for the number of people active in the scene was moderate (CCC = 0.55), and MAE was 2.57 people (64.3% of the ground truth median). Agreement between EVIP and ground truth was somewhat improved when values were aggregated across 30-s clips, particularly for number active, for which the CCC went from moderate to good.

Figure 2A (left) shows that, for number of people in the scene, EVIP errors tended to be largest at very low (<4) and very high (>46) ground truth values. These plots also show that the variability in errors tended to be larger for ground truth values that were represented in a smaller number of clips. The outliers reflected in the boxplot for ground truth values of 38 and 39 were traced to EVIP measurements from two videos.

Figure 2B (right) shows that, for number of people active, EVIP error was generally small when there were between 0 and 10 people active based on ground truth, and then increased in the direction of underestimation as the true number of people active increased. Virtually all EVIP estimates (99.8%) of the number of people active were ≤12 people, and none were >14, whereas 6.1% of ground truth values were >14. The EVIP underestimation coincided with a reduction in the number of clips available, such that the number of seconds representing each ground truth value generally decreased as the number of people active increased.

Figure 3 shows that the EVIP error was uncorrelated with EVIP estimates/predictions. Furthermore, EVIP error for the number of people in the scene and the number of people active was not associated with camera placement or the percentage of

TABLE 1. Performance of the EVIP algorithms as compared with ground truth.

Data Level and Target Variable	Median (IQR; Range)		Mean Absolute Error (95% CI)	CCC (95% CI)
	Ground Truth	EVIP		
1-s clips (N = 38,658)				
No. people in scene	16 (11–23; 0–50)	16 (11–21; 2–47)	2.70 (1.87, 3.51)	0.88 (0.74, 0.96)
Number active in scene	4 (2–8; 0–32)	4 (2–7; 0–14)	2.57 (1.94, 3.00)	0.55 (0.40, 0.70)
30-s clips (N = 1254)[a]				
No. people in scene	16.4 (11.3–22.8; 0–50.0)	15.7 (11.5–21.0; 2.4–45.1)	2.56 (1.79, 3.29)	0.89 (0.75, 0.96)
Number active in scene	4.1 (2.0–7.7; 0–20.0)	4.4 (2.3–7.0; 0.8–12.6)	2.20 (1.63, 2.60)	0.62 (0.48, 0.75)

[a]Second-level EVIP and ground truth values were averaged across each 30-s period.
Ground truth was captured using video observations (people counts) and accelerometers (activity level).
CI = confidence interval.

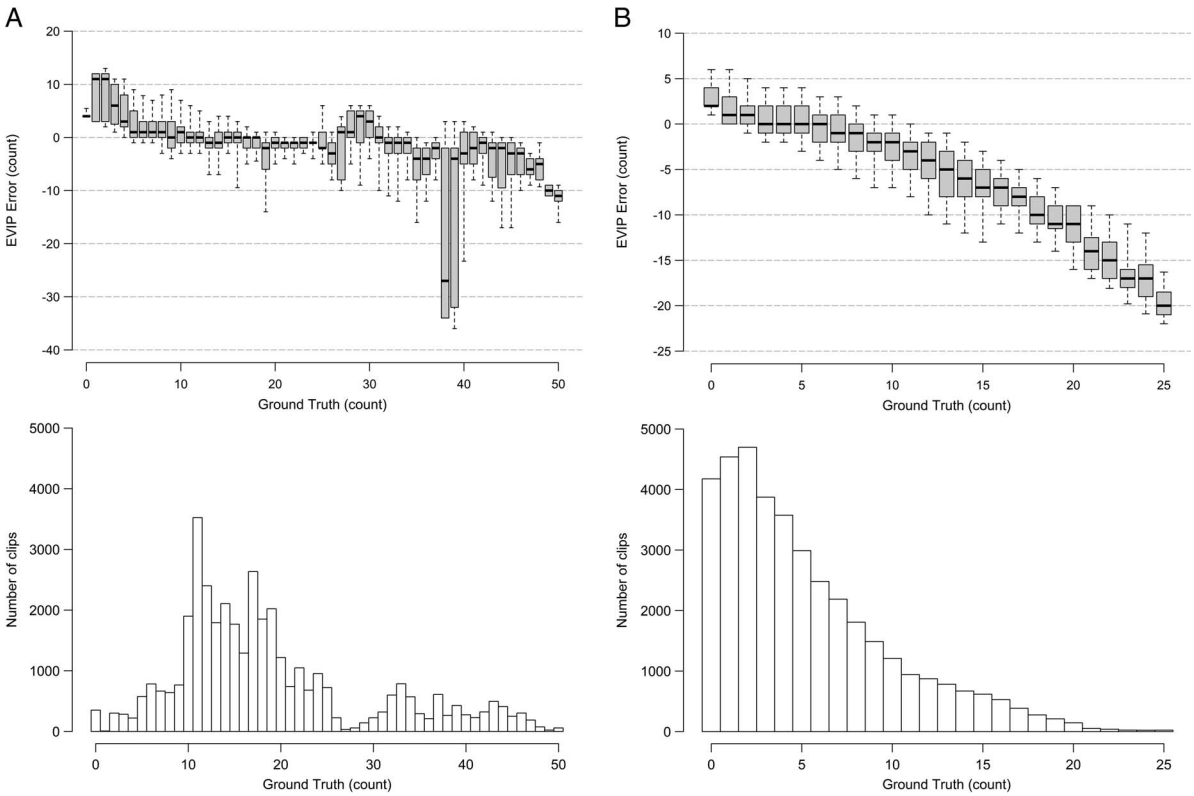

FIGURE 2—Error boxplots (top) and number of testing clips (bottom) for number of people in scene and number active Note: The top plots show EVIP error predicted − ground truth) in the number of total people (A) and number active (B) as a function of the ground truth values for number of total people and number active; whiskers indicate 5th and 95th percentiles of observed error distribution; the bottom plots show the distribution in the number of total people (A) and number active (B) across video clips in the testing data set.

the scene obstructed or shadowed, and errors related to feature type were 1.5 people or smaller (Table 2).

For both number of people in the scene and number active, EVIP outperformed SOPARC observation of the video data in estimating ground truth values (Table 3). For number of people in scene, the EVIP MAE was 41% smaller than the SOPARC MAE (2.50 vs 4.24), and the EVIP CCC was higher by 0.16 (0.88 vs 0.72). For number of people active, the EVIP MAE was 48% smaller than the direct observation MAE (2.14 vs 4.11), and the EVIP CCC was higher by 0.11 (0.66 vs 0.55).

DISCUSSION

Present findings indicated the computer vision–based EVIP algorithms can validly estimate the number of people in target areas captured by overhead video in settings such as parks and schoolyards. EVIP's validity for estimating the number of people active was less strong but considered moderate-to-good according to established criteria for interpreting CCC, and its validity was more favorable than that of traditional SOPARC observation methods. Though more work is needed to improve the algorithms, particularly for estimating the number of people who are active in more people-dense scenes, evidence is accumulating in support of the use of computer vision in physical activity research and practice (15,23).

Substantial work has been conducted in the field of computer vision related to counting people in crowded areas (24,25). However, the present study is among the first to train and show validity of a computer vision algorithm for counting

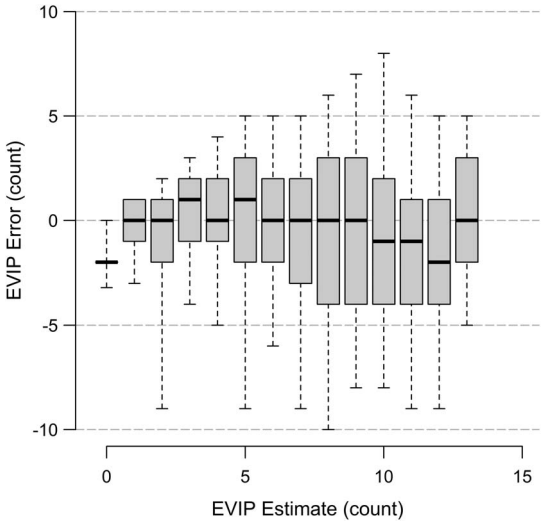

FIGURE 3—Error boxplots by EVIP predicted number active in scene Notes: The plot shows EVIP error (predicted − ground truth) in the number active as a function of the predicted values for number of people active; whiskers indicate 5th and 95th percentiles of observed error distribution; predicted counts ≥14 (n = 2) excluded.

TABLE 2. Associations of site and scene features with EVIP error (N = 38,658).

	EVIP Error (Predicted − Ground Truth), Estimate (95% CI)[a]	
	No. People in Scene	Number Active in Scene
Model 1		
Camera height (ft)	0.0 (–0.2, 0.2)	0.0 (–0.2, 0.2)
Camera distance (ft)	0.1 (–0.1, 0.3)	0.0 (–0.2, 0.2)
Percent of scene obstructed[b]	0.2 (–1.0, 1.0)	–0.1 (–0.7, 0.7)
Percent of scene covered by shadows[b]	0.1 (–0.4, 0.2)	0.0 (–0.2, 0.3)
Model 2		
Feature type		
Open green space/sports field	–0.8 (–2.2, 0.4)	–0.4 (–1.4, 0.5)
Gymnasium	–1.4 (–3.4, 0.6)	1.5 (–5.3, 7.0)
Paved surface	–1.1 (–3.1, 0.9)	–0.4 (–2.1, 1.1)
Playground	0.3 (–0.7, 1.4)	0.2 (–0.9, 1.1)

[a]95% bootstrap confidence interval.
[b]Scaled: 1 unit change = 10 percentage point change.
Ground truth was captured using video observations (people counts) and accelerometers (activity level).

people in organized physical activity settings for public health research. The people counting feature of EVIP alone, without estimating number active, could make a novel contribution to public health research and practice. Such information would inform decision making in health-related settings such as parks, for which leaders (e.g., park managers) commonly have little-to-no information on the number of people who use the park and areas of the park that are most used. Given the strong validity observed in the present study for estimating number of people, next steps should be to explore the use of EVIP and similar tools for supporting decision making in such settings.

EVIP's most significant and novel contribution is its estimation of the number of people in the scene who are active. Validity estimates were only moderate-to-good and distinguishing MVPA from sedentary/light activity was a challenging task for the algorithms. This was apparent given the large variations in error across clips and MAE that suggested an average over- or under-estimation of 2.2 people (≈54%) in physical activity in the 30-s clips. More work is needed to train more robust algorithms that can consistently estimate with accuracy and provide more valid estimates in circumstances involving large numbers of people.

In regard to understanding situations when EVIP fails, there was evidence that error increased as the number active increased, which was likely at least partly due to the underrepresentation in the data set of scenes with many people active. Although it may appear that the algorithm could be improved by adjusting estimates upward as the estimated number of people active increases, this is not likely because there was little evidence that EVIP errors were negatively correlated with EVIP estimates themselves. Future studies should strive to include more of these rarer but potentially more challenging people-dense and highly active scenes. Other potential sources of error that were measured did not appear to impact error, including occlusions rated at the level of the video/perspective and setting type. Qualitative review of video suggested that EVIP had the lowest accuracy in the sites in which people were smallest/furthest from the camera. Multiscale problems (in the case of EVIP having both large/near and small/far people) are well known in computer vision. Although researchers have created tools for overcoming multiscale problems (26), more work is need to create tools that can address the particularly small size of people in many EVIP scenes.

EVIP addresses many of the limitations of traditional SOPARC observations through automation and high-frequency data capture. Traditional SOPARC observation scans are known to involve error (11), so it was important to investigate how EVIP error compared with error involved in SOPARC observations. The finding that error was slightly lower for EVIP than SOPARC observations provides additional evidence in support of EVIP. It is important to note that the video was not stopped or rewound during the SOPARC observations, which would have likely led to increased accuracy. It is also important to note that SOPARC observations were not made in real-time in this study but on the video data, likely improving observation accuracy over field-based observations, which are subject to greater distraction. A major source of error in human observation relates to sampling and the use of momentary assessment, which involves observing each person one at a time only for a moment (4,8). EVIP overcomes these limitations by capturing every person every second – a primary advantage of the high-frequency data capture made possible by automation. High-frequency data capture also supports the capture of short-term variations in setting-based activity, such as comparing the number active across the duration of a 30-min activity (e.g., sports practice) (27,28). Similarly, automation supports researchers and practitioners to better capture a representation of activity in a setting by assessing multiple full days of data rather than relying on a sample of brief time periods across a limited number of days, which is the standard of practice for field-based SOPARC observations (8). Another unique advantage of automated assessment is that it has potential for supporting just-in-time monitoring and feedback systems to inform interventions (12,13,29).

TABLE 3. Performance of EVIP and SOPARC observations in predicting ground truth values (N = 130 observations).

Target Variable and Data Source	Median (IQR; Range)	Mean Absolute Error (95% CI)	CCC (95% CI)
No. people in scene			
Ground truth	16.0 (11.2–22.0; 1.9–46.8)	—	—
EVIP	15.8 (11.3–21.0; 4.5–43.2)	2.50 (1.87, 3.22)	0.88 (0.77, 0.95)
SOPARC observations	14.5 (10.3, 21.0; 0–36)	4.24 (3.36, 5.17)	0.72 (0.64, 0.79)
Number active in scene			
Ground truth	3.5 (1.9–7.6; 0.0–18.2)	—	—
EVIP	4.1 (2.0–7.5; 0.8–11.6)	2.14 (1.78, 2.55)	0.66 (0.58, 0.76)
SOPARC observations	8.0 (5.0, 11.8; 0–25)	4.11 (3.57, 4.71)	0.55 (0.42, 0.65)

Ground truth was captured using video observations (people counts) and accelerometers (activity level); second-level ground truth and EVIP values were averaged across the duration of each SOPARC observation period.

Strengths, limitations, and future directions. Strengths of the present study included capturing a variety of setting types, free-living activities, and people densities; investigating potential sources of bias; and comparing performance between EVIP and traditional SOPARC observations. There were several limitations that should be considered when interpreting the findings. The EVIP algorithms were both trained and tested on all sites, so the validity of EVIP when applied to a new untrained site is unknown. Calibration data and processes are likely to be needed to "transfer" the algorithms to a new site. A larger number of sites with greater diversity is likely needed to train algorithms that can generalize across sites and should be captured in future studies. All data were collected during daylight, so the algorithms are not likely to perform well in the dark. The ground truth measure was limited by providing only scene-level information (the total number of people and number active). Using person-/location-linked ground truth information, which could be done with high quality using direct observation of video, would allow the use of additional computer vision modules and techniques and should be explored in future studies. Rigorous second-by-second video observations may also provide a more accurate ground truth measure than that used in the present study. EVIP is currently not readily usable without computer vision expertise. Future work is needed to package and disseminate validated algorithms in a user-friendly platform. A limitation of the camera approach compared with human in-person observations is the restricted field of view. A limitation to automated video assessment is the inability to capture equity-related variables such as age, sex, and race/ethnicity, though these characteristics are also challenging to collect via human observations.

CONCLUSIONS

The present study found that novel computer vision algorithms can be used to estimate the number of total people and number people active from video recordings of target areas in parks and schoolyards with moderate-to-good validity, when trained on the same settings they are applied to. Thus, computer vision appears promising for automating ecological assessment of setting-based physical activity, but further work is needed to create generalizable algorithms with low error variability. Such automated high-frequency data capture is advantageous because it can circumvent some of the key limitations to traditional field-based direct observations and support just-in-time monitoring and feedback for those delivering interventions. More research is needed on the application of computer vision-based physical activity assessment tools in both observational and intervention studies in a variety of settings.

Funding was provided by NIH grant R21CA194492. Thank you to Alexia Jadlow, Arwen Marker, Ashleigh Galler, Carolina Bejarano, Chelsea Steel, Farhia Jeilani, Jack Sanchez, Jacqueline Pierre, Jonathan Finch, Kelli Snow, Laura Graber, Matilian Cassmeyer, and Sean Wheaton for their work on this study. The results of this study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results do not constitute endorsement by ACSM. The authors report no conflicts of interest.

REFERENCES

1. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. *Annu Rev Public Health*. 2006;27:297–322.
2. McKenzie TL, Sallis JF, Nader PR. SOFIT: system for observing fitness instruction time. *J Teach Phys Educ*. 1992;11(2):195–205.
3. McKenzie TL. 2009 C. H. McCloy Lecture. Seeing is believing: observing physical activity and its contexts. *Res Q Exerc Sport*. 2010; 81(2):113–22.
4. McKenzie TL, Van Der Mars H. Top 10 research questions related to assessing physical activity and its contexts using systematic observation. *Res Q Exerc Sport*. 2015;86(1):13–29.
5. Cohen DA, Han B, Isacoff J, et al. Impact of park renovations on park use and park-based physical activity. *J Phys Act Health*. 2015;12(2): 289–95.
6. Cohen DA, Han B, Nagel CJ, et al. The first national study of neighborhood parks: implications for physical activity. *Am J Prev Med*. 2016;51(4):419–26.
7. Hollis JL, Williams AJ, Sutherland R, et al. A systematic review and meta-analysis of moderate-to-vigorous physical activity levels in elementary school physical education lessons. *Prev Med*. 2016;86:34–54.
8. McKenzie TL, Cohen D. *SOPARC (System for observing play and recreation in communities): Description and Procedures Manual*. San Diego (CA): San Diego State University; 2006.
9. Cohen DA, Setodji C, Evenson KR, et al. How much observation is enough? Refining the administration of SOPARC. *J Phys Act Health*. 2011;8(8):1117–23.
10. McKenzie TL, Cohen DA, Sehgal A, Williamson S, Golinelli D. System for Observing Play and Recreation in Communities (SOPARC): reliability and feasibility measures. *J Phys Act Health*. 2006;3(1 Suppl): S208–22.
11. Weaver RG, Beighle A, Erwin H, Whitfield M, Beets MW, Hardin JW. Identifying and quantifying the unintended variability in common systematic observation instruments to measure youth physical activity. *J Phys Act Health*. 2018;15(9):651–60.
12. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. 2017;52(6):446–62.
13. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. *Health Psychol*. 2015;34S:1209–19.
14. Cyganek B, Siebert JP. An introduction to 3D computer vision techniques and algorithms: John Wiley & Sons; 2011.
15. Carlson JA, Liu B, Sallis JF, et al. Automated ecological assessment of physical activity: advancing direct observation. *Int J Environ Res Public Health*. 2017;14(12):E1487.
16. Freedson P, Pober D, Janz KF. Calibration of accelerometer output for children. *Med Sci Sports Exerc*. 2005;37:S523–30.
17. Trost SG, Loprinzi PD, Moore R, Pfeiffer KA. Comparison of accelerometer cut points for predicting activity intensity in youth. *Med Sci Sports Exerc*. 2011;43(7):1360–8.
18. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In: *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition*; 2014: Columbus, Ohio (USA). p. 1725–1732.

19. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. Santiago (Chile); 2015. pp. 4489–97.
20. Lin L, Hedayat A, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues, and tools. *J Am Stat Assoc*. 2002;97(457): 257–70.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74.
22. The R Foundation for Statistical Computing. The R Project for Statistical Computing, 2019 [Oct 16, 2019]. Available from: https://www.r-project.org/.
23. Silva P, Santiago C, Reis LP, Sousa A, Mota J, Welk G. Assessing physical activity intensity by video analysis. *Physiol Meas*. 2015; 36(5):1037–46.
24. Zhang Y, Zhou D, Chen S, Gao S, Ma Y. Single-image crowd counting via multi-column convolutional neural network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (USA); 2016. pp. 589–97.
25. Shi M, Yang Z, Xu C, Chen Q. Revisiting perspective information for efficient crowd counting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach, CA (USA); 2019. pp. 7279–88.
26. Cai Z, Fan Q, Feris RS, Vasconcelos N. A unified multi-scale deep convolutional neural network for fast object detection. In: *European conference on computer vision*. Amterdam (Netherlands); 2016. pp. 354–70.
27. Schlechter CR, Guagliano JM, Rosenkranz RR, Milliken GA, Dzewaltowski DA. Physical activity patterns across time-segmented youth sport flag football practice. *BMC Public Health*. 2018;18(1):226.
28. Schlechter CR, Rosenkranz RR, Fees BS, Dzewaltowski DA. Preschool daily patterns of physical activity driven by location and social context. *J Sch Health*. 2017;87(3):194–9.
29. Trowbridge MJ, Huang TT, Botchwey ND, et al. Public health and the green building industry: partnership opportunities for childhood obesity prevention. *Am J Prev Med*. 2013;44(5): 489–95.