International Journal of Semantic Computing
Vol. 11, No. 2 (2017) 1–34
©World Scientific Publishing Company
DOI: 10.1142/S1793351X17002751

World Scientific
www.worldscientific.com

Enriching the Fan Experience in a Smart Stadium Using Internet of Things Technologies

Sethuraman Panchanathan*, Shayok Chakraborty†, Troy Mcdaniel‡, Ramin Tadayon§ and Bijan Fakhri¶

Center for Cognitive Ubiquitous Computing (CUbiC)
Arizona State University, Tempe, AZ 85281, USA
*panch@asu.edu
†shayok.chakraborty@asu.edu
‡troy.mcdaniel@asu.edu
§rtadayon@asu.edu
¶bfakhri@asu.edu
http://cubic.asu.edu

Noel O'Connor‖, Mark Marsden**, Suzanne Little†† and Kevin Mcguinness‡‡

Insight Centre for Data Analysis, Dublin City University
Glasnevin, Dublin 9, Ireland
‖noel.oconnor@dcu.ie
**mark.marsden@insight-centre.org
††suzanne.little@dcu.ie
‡‡kevin.mcguinness@dcu.ie

David Monaghan

School of Computer Science and Statistics
Trinity College Dublin, College Green, Dublin 2, Ireland
monaghd2@tcd.ie

Rapid urbanization has brought about an influx of people to cities, tipping the scale between urban and rural living. Population predictions estimate that 64% of the global population will reside in cities by 2050. To meet the growing resource needs, improve management, reduce complexities, and eliminate unnecessary costs while enhancing the quality of life of citizens, cities are increasingly exploring open innovation frameworks and smart city initiatives that target priority areas including transportation, sustainability, and security. The size and heterogeneity of urban centers impede progress of technological innovations for smart cities. We propose a Smart Stadium as a living laboratory to balance both size and heterogeneity so that smart city solutions and Internet of Things (IoT) technologies may be deployed and tested within an environment small enough to practically trial but large and diverse enough to evaluate scalability and efficacy. The Smart Stadium for Smarter Living initiative brings together multiple institutions and partners including Arizona State University (ASU), Dublin City University (DCU), Intel Corporation, and Gaelic Athletic Association (GAA), to turn ASU's Sun Devil Stadium and Ireland's Croke Park Stadium into twinned smart stadia to investigate IoT and smart city technologies and applications.

May 25, 2017 6:11:43pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

Keywords: Internet of things; smart stadium; smart city; crowd behavior analytics; object counting.

1. Introduction

People increasingly moving to urban centers is shifting the balance between rural and city life. This phenomenon of rapid urbanization has brought about significant changes in where the global population resides: 54% of the global population was in urban in 2014, and by 2050, estimates predict that 64% of the global population will be urban [41]. Rapid urbanization is exacerbating existing concerns of congestion, pollution, accidents, security, and sustainability. For example, it is estimated that by 2050, the number of vehicles on the road will double to 2.5 billion. In 2013, the U.S. spent $124 billion due to traffic congestion, and estimates predict that by 2030, this number will rise to $186 billion with accompanying increases in "social costs" [59]. By 2020, $13 billion and 1,600 premature deaths are anticipated in costs due to exposure to emissions from idling vehicles during traffic jams. Traffic congestion problems are a worldwide issue; as of 2014, the top 10 most congested cities [59] include Istanbul, Mexico City, Rio de Janeiro, Moscow, Salvadore, Recife, St. Petersburg, Bucharest, Warsaw, and Los Angeles.

Cities are seeking ways to reduce complexity and costs, provide better management, and meet resource needs, while ensuring a high quality of life for its citizens. Many cities have begun to explore open innovation frameworks and smart city initiatives to address the needs of their growing populaces by targeting key priority areas of health, wellness, transportation, safety, security, sustainability, and citizen engagement. Cities that perform well and excel will flourish through the creation of wealth and rises in productivity, paving the way for continued growth and long-term success [29]. Smart city transformations rely upon not only technological and policy-based advancements, but re-imagining traditional approaches to key priority areas, and preparing for scalability challenges due to a city's sheer size and heterogeneity. We propose the use of a Smart Stadium as a living laboratory to more easily deploy and evaluate Internet of Things (IoT) technologies and smart city solutions by balancing the size and heterogeneity of a smart environment that is small enough to practically trial but large and complex enough to evaluate effectiveness and scalability.

Smart Stadium for Smart Living is an initiative developed to join institutions and partners interested in IoT and smart city technologies. The initiative joins Arizona State University (ASU) in Tempe, Arizona; Dublin City University (DCU) in Dublin, Ireland; Gaelic Athletic Association (GAA) of Ireland; and Intel Corporation to turn two stadia — ASU's Sun Devil Stadium and Ireland's Croke Park Stadium — into twinned smart stadia with the potential to be world class testbeds for exploring smart city applications and IoT solutions. The projects of this initiative thus far focus on two broad application areas: (i) Enriching the fan/attendee experience; and (ii) Enhancing stadium operation. While the application focus of these

May 25, 2017 6:11:43pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

projects is set in the context of the stadium and stadium-related events, they are relevant to wider smart city application areas. The full scope of projects within this initiative addresses issues of crowd management, fan engagement, event logistics, stadium management, and environmental monitoring, using a variety of deployed sensors such as video cameras and microphones. Given the sheer number of projects within this initiative, the following discussion pertains only to projects targeting enriching the fan experience.

Projects to enrich a fan's experience were identified by considering the entire 'journey' of an event attendee; that is, not only his or her interactions, behaviors, and actions within the stadium, but all activities involved to attend an event. For example, a fan's journey may include extensive preparation, perhaps months prior, to attend an upcoming event; planning and coordination to travel to and from the stadium; their involvement on social media leading up to an event as well as during and after an event itself; and activities carrying over to relevant events and gatherings happening before, during, and after the stadium event itself. This work presents three fan-focused projects targeting efficiency/convenience, safety, and engagement. These projects include: (i) *Crowd Understanding*: Improved safety via vision-based and non-vision-based crowd behavior understanding and analytics; (ii) *Athletic Demonstrator Platform*: Interactive serious gaming stations to support fan engagement while promoting motor learning and athletic training; and (iii) *Wait Time and Queue Estimation*: Real-time, accurate access via a mobile app to wait time estimates of lines across a stadium's concession stands, souvenir stands, and restrooms.

1.1. *Organization and research contributions*

The rest of this paper is organized as follows: Section 2 discusses the *Crowd Understanding* project. We present an efficient strategy to compute low dimensional, informative features for crowd behavior understanding and anomaly detection.

The *Athletic Demonstrator Platform* project, outlined in Sec. 3, is a motor learning environment enabling real-time motion capture, analysis, and feedback. The main contributions of this work include: (1) A fusion approach for low-cost Kinect-IMU motion capture and algorithms for calibration, phase detection, and analysis; and (2) Insight into important research questions pertaining to the design of multimodal feedback including (i) What categories of performance are present in real motor training feedback from a trainer to a subject, and through which modalities do these interactions occur? (ii) How can a system observe these metrics of performance in an individual's motion? and (iii) Does individual preference play a role in the assignment of modalities to feedback in a multimodal environment?

Section 4 presents the *Wait Time and Queue Estimation* project. Our research contributions in this project include: (1) A novel active learning framework to identify the salient and exemplar instances from large amounts of unlabeled data to train an object counting model and (2) Incorporating only binary (yes/no) feedback into the algorithm in order to reduce the labeling burden on the user.

Finally, we conclude with discussions in Sec. 5.

2. Crowd Understanding

Sports Stadiums are multi-purpose venues within our cities where thousands gather for events including sporting contests, music concerts as well as business and academic conferences. However, with such large gatherings of people there are significant risks to public safety which must be addressed. Improving our understanding of the behavior of such large crowds of people within a stadium can help maintain safety and security for all involved. Early detection and a rapid response time are essential in any emergency situation, especially in a highly congested public space such as a stadium. To address this issue, we have developed an efficient computer vision algorithm for detecting unusual crowd behavior in real-time on a commodity CPU. Both Sun Devil Stadium (56,200 capacity) and Croke Park Stadium (82,300 capacity) have been fully designed to ensure the safety of all visitors, but the Smart Stadium project aims to exploit visual and non-visual sensor data to gain additional insight into the dynamics of crowds which will help improve the already excellent safety standards.

The crowd understanding project uses existing CCTV camera footage from Croke Park Stadium to extract scene-level holistic features and detect unusual crowd behavior at the frame level. Long-term, the system aims to learn a "steady state" of what normal crowd behavior patterns look like across numerous cameras within a stadium, and therefore, be able to determine when crowds don't behave according to expected patterns, and alert support staff.

2.1. *Crowd understanding implementation*

The objective is to design a low dimensional set of features that are quick to compute and capture sufficient holistic information about objects moving in a scene to allow straightforward discrimination between normal and abnormal events. The developed technique for crowd behavior anomaly detection uses a set of efficiently computed, easily interpretable, scene-level holistic features [40]. These features are calculated by analyzing local motion patterns across a crowded scene. This low-dimensional descriptor combines two features from the literature: crowd collectiveness [52] and crowd conflict [27], with two newly developed features: mean motion speed and a unique formulation of crowd density [40].

Crowd collectiveness is a scene-independent holistic property of a crowd system, which can be defined as the degree to which individuals in a scene move in unison [52]. Zhou *et al.*'s [52] method for measuring this property analyzes the tracklet positions and velocities found in the current frame and constructs a weighted adjacency matrix. The edge weights within each matrix column are summed and the

May 25, 2017 6:11:43pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

mean is calculated. This mean value corresponds to the overall collectiveness level for the current frame.

Crowd conflict is another scene-independent holistic crowd property, which can be defined as the level of friction/interaction between neighboring tracked points [52]. Shao *et al.* [52] efficiently calculate this property by summing the velocity correlation between each pair of neighboring tracked points in a given frame.

Crowd density can be defined as the level of congestion observed across a scene at a given instant. The proposed approach to calculating this feature firstly divides the scene into a fixed size grid (10×10) and counts the number of grid cells currently occupied by one or more tracked points. Equation (1) is then used to calculate the crowd density level for the current frame. A 10×10 grid was chosen to provide sufficient granularity in the density calculation, with the aim being for each grid cell to roughly contain one or two pedestrians in most surveillance scenarios. There are obvious limitations in terms of scale invariance with this feature, however the main objective is not pixel perfect accuracy but to measure a useful crowd property in a highly efficient manner.

$$\text{CrowdDensity} = \frac{\text{Occupied Grid Cells}}{\text{Total Grid Cells}} \tag{1}$$

Figure 1 depicts the proposed crowd density feature calculated using footage from a CCTV camera covering a concession area at Croke Park Stadium during a busy match. As shown, the density level increases significantly once the gates open (yellow), fall once the match begins (green), and then spike again at half time (red).

The heat map in Fig. 1 is taken from a video animation that illustrates how the density level at various stadium locations changes over time. This was produced by calculating the density level over a full match day for each camera location and updating the color for each section to correspond to the density level [0.0–1.0] at that time point. Using this method, the distribution of people throughout a stadium over the course of a match day can be visualized. The visualization can also be sped up to show the changes the take place over hours in a matter of minutes. Figure 2 shows this feature being calculated on an image from the UMN dataset.

The mean motion speed observed within a crowded scene provides a coarse, scene-level feature that can be extracted very efficiently. Our approach estimates this crowd property by calculating the magnitude of each tracklet velocity vector in the current frame and finding the mean. While conceptually simple, experiments show that the inclusion of this feature noticeably improves the accuracy of crowd behavior anomaly detection. Each of these features captures a distinct aspect of crowd behavior.

Our holistic features are extracted for each frame in a given video sequence using the following steps. Firstly, the scene foreground is segmented using the Gaussian-mixture based method of KaewTraKulPong and Bowden [33] before interest points are tracked using a KLT tracker [58]. These local trajectories or tracklets are then analyzed to calculate four holistic features for each frame. This high-level descriptor

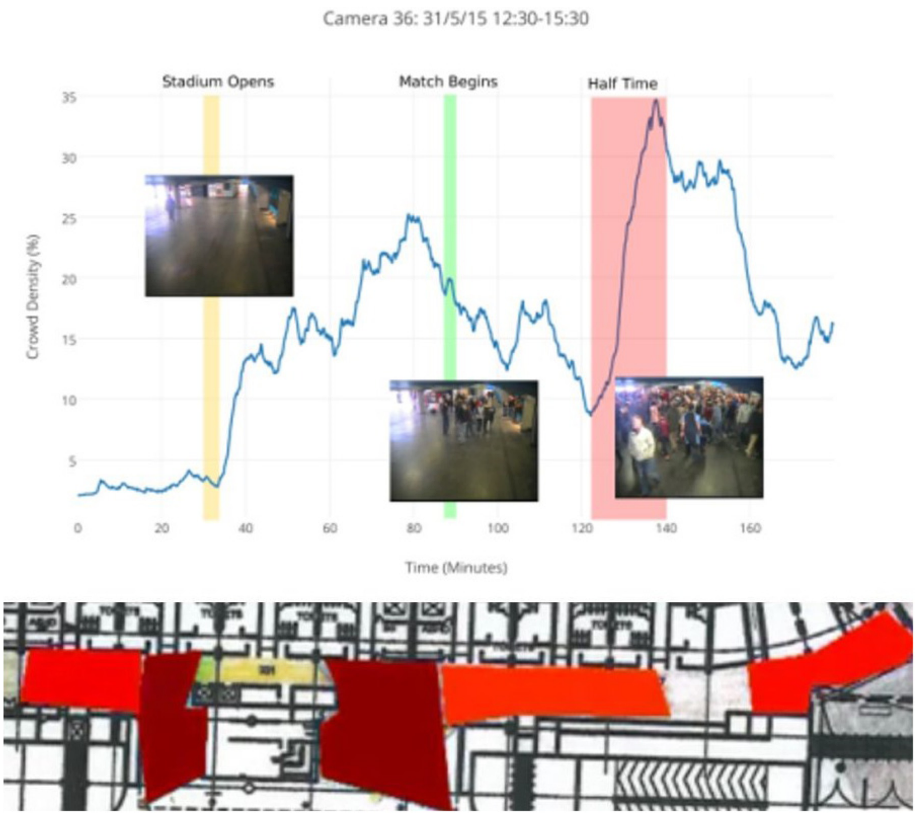


Fig. 1. (Color online) *Top*: Changes in crowd density calculated for a concession stand during a busy match day at Croke Park Stadium. *Bottom*: A heat map visualization showing differences in crowd density at different stadium locations within Croke Park.

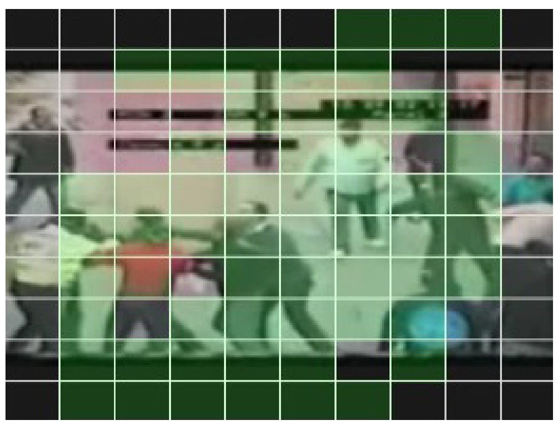


Fig. 2. Crowd density calculation grid for a scene from the violent-flows dataset. Each green square corresponds to an occupied grid cell (crowd density in this frame = 57%).

of crowd behavior can be computed in real-time (30+ frames per second) even on commodity hardware (e.g., an Intel i5 CPU).

Anomalous crowd behavior then needs to be detected using this crowd behavior descriptor. We investigate two anomaly detection approaches, covering two possible situations: (1) When only "Normal" behavior training data is available; and (2) when both "Normal" and "Abnormal" behavior training data are available. Each require the following pre-processing steps: All individual features are firstly scaled to lie within the range $[0, 1]$, with respect to the range of training data values. Normalization is then performed by dividing by the maximum magnitude vector in the training set. The low-dimensional descriptor used results in almost negligible training and classification times for reasonably sized datasets.

We use a Gaussian Mixture Model (GMM) to perform outlier detection when only normal behavior training data is available. The GMM configuration (number of mixture components and type of co-variance matrix) for a given experiment is selected as the one that minimizes the Bayesian Information Criterion (BIC) value [48] on the training data. The selected model is then used to calculate the log probabilities for the full set of training frames, and the distribution of these log probability values is used to decide upon an outlier detection threshold using Otsu's method [43]. Test frames are then classified as abnormal or normal by using the fit mixture model to calculate their log probability and applying the adaptive threshold generated from the training data.

We use a discriminative model (binary classifier) for outlier detection when both normal and abnormal training data are available. Specifically, we trained a Support Vector Machine (SVM) with an RBF kernel on test frames labeled as normal and abnormal. The default value of 1.0 was used for the SVM regularization parameter C.

2.2. *Crowd understanding results*

The proposed method is evaluated on two distinct crowd behavior anomaly datasets: (i) the UMN dataset[a]; and (ii) the violent-flows dataset [27]. These benchmarks assess the ability of a given approach to detect unusual crowd behavior at the frame-level and video-level, respectively. All experiments were carried out using MATLAB 2014a and Python 2.7 on a 2.8 GHz Intel Core i5 processor with 8GB of RAM.

The UMN dataset contains 11 sequences filmed in 3 different locations. Each sequence begins with a period of normal passive behavior before a panic event/anomaly occurs toward the end. The objective here is to train a classifier using frames from the initial normal period and evaluate its detection performance on the subsequent test frames. Classification is performed at the frame level and results are compared in terms of the receiver operating characteristic (ROC) curve's area under the curve (AUC). For each of the three scenes, the initial 200 frames of each clip are combined to form a training set, with the remaining frames used as a test set for that

[a] http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.

Table 1. BIC values calculated during the GMM selection stage for the UMN dataset.

No. of mixture components	BIC
1	−20015
2	−21810
3	**−22047**
4	−21940

scene. This results in a roughly 1:2 split between training and test frames for each camera location and will be referred to as the single scene experiment. While this dataset is quite limited in terms of size and variation, it does provide a good means of performance evaluation during the development of a crowd anomaly detection algorithm. Since no abnormal frames are made available for training in this experiment, the GMM-based detection approach is used. Table 1 presents the BIC values calculated during the GMM selection stage, with a 3-component model ultimately used. A full co-variance matrix GMM resulted in a lower BIC value in all cases and was therefore used. Figure 3 presents the ROC curves for all three UMN scenes individually. A cross-scene anomaly detection approach is also taken, where for a

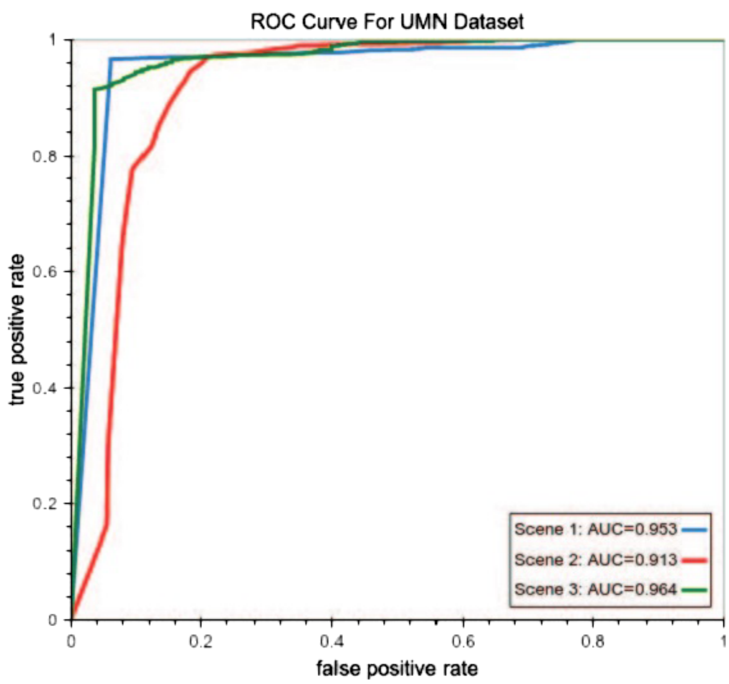


Fig. 3. Receiver operating characteristic (ROC) curve and associated area under the curve (AUC) for each UMN scene.

May 25, 2017 6:11:49pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

Table 2. ROC curve AUC performance and processing speed on the UMN dataset.

Method	AUC	Speed (FPS)
MDT	0.995	0.9
CM	**0.98**	5
SFM	0.97	3
Proposed Method (Single Scene)	0.929	**40**
Proposed Method (Cross-Scene)	0.869	**40**

given UMN scene, the training frames from the two other scenes are used to generate the GMM.

Table 2 compares the two variants of the proposed method with the leading approaches in terms of AUC and processing frame rate. The proposed approach achieves competitive classification performance with the state-of-the-art at just a fraction of the computational cost. The cross-scene experiment, while inferior in terms of classification performance, is noteworthy in that each scene was classified using training data only from other surveillance scenarios.

The violent-flows dataset contains 246 clips containing violent (abnormal) and non-violent crowd behavior. Classification is performed at the video level. A 5-fold cross validation evaluation approach is taken and results are compared in terms of mean accuracy. As both normal and abnormal training examples are available in this dataset, the proposed SVM-based classification approach is used. The majority classification found among the frames of a given clip is used as the overall result for that clip. An alternate approach is also taken where only the normal training examples are used, and the proposed GMM-based outlier detection approach is taken.

Table 3 presents the BIC values calculated during the GMM selection stage, with a 4-component model ultimately used. A full co-variance matrix GMM resulted in a lower BIC value in all cases and was therefore used. For this GMM-based approach the histogram of frame log probabilities for a given test clip is generated and the mode value is used to classify the overall clip by applying the Otsu threshold generated from the training data. Table 4 compares the two variations of the proposed technique with the leading approaches in terms of mean accuracy and processing

Table 3. BIC values calculated during the GMM selection stage for the violent-flows dataset.

No. of mixture components	BIC
1	−51758
2	−223161
3	−274742
4	**−327545**

Table 4. Mean accuracy and processing speed on the violent-flows dataset.

Method	Accuracy (%)	Speed (FPS)
SD	85.4	N/A
HOT	82.3	N/A
ViF	81.3	30
CM	81.5	5
Proposed Method (SVM)	**85.53**	**40**
Proposed Method (GMM)	65.8	**40**

Table 5. The contribution of each feature toward mean detection accuracy on the violent-flows dataset using proposed SVM-based detection approach.

Feature	Accuracy when excluded (%)
Crowd Collectiveness	75.2
Crowd Conflict	65.5
Crowd Density	63.5
Mean Motion Speed	81.2

frame rate. Table 5 highlights the contribution of each feature towards the achieved anomaly detection accuracy on the violent-flows dataset using the SVM-based variant. As shown, leaving out any individual feature results in a noticeable decrease in anomaly detection accuracy.

The SVM-based variant achieves state-of-the-art performance on the violent-flows dataset with a mean accuracy of $85.53 \pm 0.17\%$. The GMM-based variant achieves a very respectable $65.8 \pm 0.15\%$ accuracy, which is particularly impressive considering only half the training data, containing no violent behavior, is used in this case. The approach also achieves noticeably faster computational performance.

The proposed scene-level holistic features are easily interpretable, sensitive to abnormal crowd behavior, and can be computed in better than real-time (40 frames per second) on commodity hardware. The approach was demonstrated to improve upon the state-of-the-art classification performance on the violent-flows dataset. Future work will attempt to improve upon certain limitations of the approach such as the scale issues present in the crowd density feature, possibly using an adaptive grid cell size. Moreover, this descriptor will be used to label specific crowd behavior concepts in larger and more challenging datasets.

3. Athletic Demonstrator Platform

Modern technology has made motion sensing more accessible and prevalent than ever before, with the rise of low-cost motion-sensing hardware such as Microsoft's Kinect camera. Similarly, multimodal feedback has become increasingly ubiquitous through the introduction of haptic, visual, and audio feedback mechanisms in phones

May 25, 2017 6:11:50pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

and game controllers, among other devices. Thanks to this evolution of technology, motor training is now more accessible to the everyday user, leading to a surge in studies on motor learning in Human-Computer Interaction (HCI). With modern technology, an automated system is capable of observing and reacting to a great deal of information pertaining to a user's motion, and with the inclusion of expert data, the system can evaluate and provide feedback on this motion in real-time, leading to a new wave of "unsupervised" motor training wherein an individual interacts with a system, rather than a real trainer, to gain proficiency in motor skills. This type of training has a variety of applications ranging from rehabilitation [54] to sports training [63]. This technology solves a critical problem in the field: a user must regularly perform and receive feedback on a motor task to improve at that task at a steady rate [11], but since trainer availability is limited, user compliance with this training can stagnate over time [46].

To provide the type of feedback on motor performance that a user can consider useful in comparison to a real trainer, an automated system should perform the following tasks: (i) The system should accurately capture a user's motion using commonly accessible technology (without the complex setup typically encountered in a laboratory or clinical environment); (ii) The system should automatically recognize, classify, and represent the various segments and elements of a motion; (iii) The system should be able to accurately interpret motion data to form an assessment of a user's performance; and (iv) The system should provide feedback on this assessment that is understandable and meaningful to the user so that the user can improve his or her motion in the next attempt.

Various aspects of the motion itself should also be considered in the provision of real-time feedback including the type of motion (rehabilitation vs. sports, for example), the user's proficiency level and previous experience, the complexity of the motion task (typically determined by observing the number of limbs involved in the motion), the type of information observed (spatial and temporal aspects of the motion), the assignment of modalities to different aspects of feedback, and the timing of feedback (for example, concurrent vs. terminal), among others.

Here we present a platform for the provision of automated multimodal feedback for motor performance in a variety of motor training scenarios. The proposed Athletic Demonstrator Platform implements real-time motion analysis and feedback to facilitate a motor learning environment that is both useful and stimulating to enrich fan engagement, excitement, and competitiveness. As part of future work, the platform has potential for athlete training.

3.1. *Related work in motion capture*

The field of Motion Capture, or "MoCap", is a widely studied area in which various techniques and methods have been applied toward the quantification and digital representation of a human's motion in an automated system [60]. Perhaps the most cutting-edge system to date for this task is the Vicon system, which uses accurate

May 25, 2017 6:11:50pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

and high-quality tracking of worn body markers to record and analyze complex motion. However, this system is expensive and often restricted to laboratory environments or professional motion capture scenarios due to its complexity, making it impractical for the typical user. As a result, cost-efficiency has become a recent concern in the field, leading to the rise of more affordable alternative systems [21, 19] which rely on computer vision [61] and depth-sensing [20, 62] to form lower-quality estimates of a user's body orientation and joint movement during a motion. Inevitably, these mechanisms are subject to the errors caused by occlusion of body parts and other issues relating to static camera sensing.

In addition to camera-based techniques, some wearable alternatives to Vicon exist for motion capture. One popular alternative is Inertial Measurement Units (IMUs), body-worn 3D motion sensors which offer an accuracy that can compete with the gold standard [2, 36]. One example of IMU application in MoCap is the XSens system [47]. These systems have seen limited success in practice due to the accumulation of calculation errors which affect the accuracy of their measurements over a time period. Furthermore, if only IMUs are used to handle motion capture, a significant amount would be needed to cover all body motion, which can be very costly. To address this issue, we can take a hybrid approach, which utilizes both worn IMUs and Kinect depth camera sensing, and fuses the readings from these two devices [18, 17] to correct for the accuracy errors of one while solving the occlusion issue of the other.

In previous work [3], we have shown that the hybrid approach, in combination of 2–3 IMU sensors with real-time joint-tracking camera data and the implementation of advanced algorithms for calibration, phase detection, and analysis, can provide a low-cost yet accurate mechanism for capture of motor activity. Here we discuss the design of a platform that utilizes the fusion approach, along with multimodal feedback in its final design, to provide learning interaction for motion of any type, depending on the location of IMU sensors worn on the body.

3.2. *Athletic demonstrator platform implementation*

The athletic demonstrator platform utilizes a combination of two IMU sensors (currently wrist-worn, but can be reconfigured), the Microsoft Kinect V2 depth camera for joint tracking, and the Unity engine for multi-platform game development, as its core components. Each IMU sensor communicates with a central computer running the platform's software over a Bluetooth connection at 256 Hz and includes accelerometer and gyroscope output for position and orientation. The system first calibrates the sensing by requiring the user to stand in a "T-pose", thus synchronizing the IMU sensors to the Kinect's coordinate space. After this, the skeletal tracking of the Kinect is fused with the readouts of the IMU sensors to determine accurate joint positioning based on techniques shown in [18, 17, 3]. The joint data tracked by movement of the worn IMUs are utilized to determine the position and orientation of those joints, while the Kinect's data is utilized to

May 25, 2017 6:11:50pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

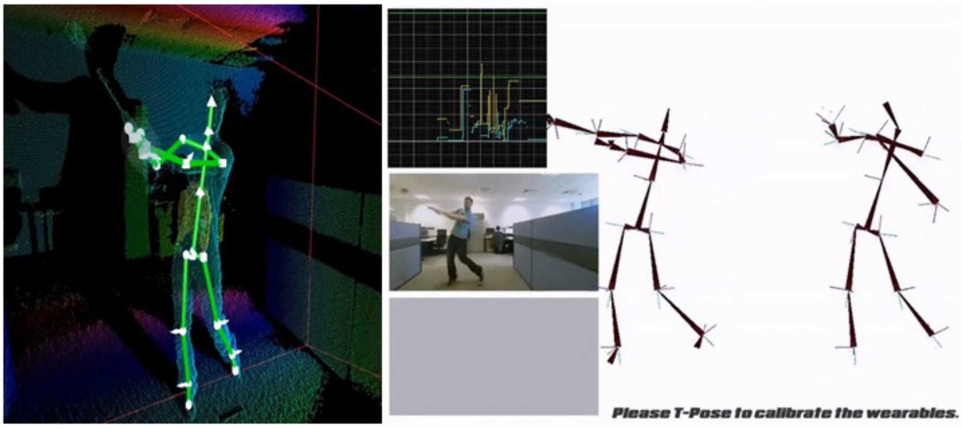


Fig. 4. Proposed low-cost Kinect-IMU motion capture fusion approach. *Left*: Demonstration of Kinect skeletal tracking. *Right*: Calibration phase and fusion of Kinect and IMU data.

determine the position and orientation of all other joints during a motion. This fusion technique is shown in Fig. 4.

To learn a motion in the current design of the platform, the user first views a demonstration of the motion by an expert through an on-screen video, which is accompanied by both an avatar representation of the fused Kinect/IMU data and a graph which depicts 3D IMU accelerometer information over time. Having viewed this demonstration, the user is then asked to attempt the motion under the same interface, with a 3D virtual avatar mirroring the user's motion as a form of concurrent visual feedback. Mechanisms are also in place for the provision of haptic feedback and audio cues at key points during this motion attempt, although the concurrent feedback used is purely visual in the initial prototype shown in Fig. 5.

Fig. 5. Athletic Demonstrator Platform. *Left*: Live demonstration of the platform for the Irish sport of hurling. *Right*: Gamified score feedback based on expert player with top ten scoreboard to promote competitiveness.

Once the user completes an attempt of the motion, he or she is then provided terminal feedback on performance using a scoring system that depicts the proximity of the user's motion, captured with both the IMU sensors and Kinect camera, to the motion sample provided by the expert. This scoring is accompanied by feedback on the user's speed, specifying whether the user should slow down or speed up the rate of motion on the next attempt. This terminal feedback provides an overview of the individual's performance for a single attempt; after several attempts, the user is given an overall score for the motion as a final measure of his or her current performance. This overall score is submitted to a leaderboard indicating the best performances on that motion, which can be used either by a single user to determine how his or her performance is progressing over time, or by multiple users to compare their performances on the same task.

3.3. *Athletic demonstrator platform: Related studies*

The athletic demonstrator platform was designed to be highly configurable, allowing for different multimodal designs for the prevision of concurrent and terminal multimodal feedback on motor performance. It was also designed to handle a large variety of motions with the fusion capture method. This flexible design has led to a series of research questions on multimodal implementation which we have addressed through research studies. These questions include: (i) What categories of performance are present in real motor training feedback from a trainer to a subject, and through which modalities do these interactions occur? (ii) How can a system observe these metrics of performance in an individual's motion? (iii) Does individual preference play a role in the assignment of modalities to feedback in a multimodal environment? Findings related to each of these questions are discussed below, and together they will inform the final design of the athletic demonstrator platform.

3.3.1. *Case study on categories of feedback*

To address the first question, real motor training scenarios were observed as part of a case study between a subject and a martial arts trainer. The goal of the first phase in this case study was to determine what forms of feedback occur in real-time as the subject interacts with the trainer, and in what modalities these interactions occur. To achieve this, a live training session was recorded between these individuals on video, and specific instances of feedback given by the trainer during the interaction were noted. For these feedback instances, both the modality of feedback and the category of feedback were determined.

Through this study, detailed in [55], three main categories of real-time motor feedback were identified: (i) Posture: a spatial measure of feedback relating to the configuration of the user's body and limbs during motion; (ii) Progression: a spatial metric which relates to the range of motion and the accuracy of an individual's motion trajectory compared to the ideal motion; and (iii) Pacing: a temporal

May 25, 2017 6:11:52pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

measure representing the speed of an individual's motion, its consistency, and its comparison to the ideal rate of motion.

Together, the three categories above constitute a complete representation of motor performance. While they were applied in this case to rehabilitative motion, these categories can be applied toward sports motions as well. For example, a football or baseball throw relies on proper configuration of the elbow and grip of the ball (posture), momentum of forward motion prior to release (pacing), and release of the ball at the correct moment to achieve an ideal trajectory (progression), as indicated in [4].

The primary modalities of feedback discovered were audio (delivered as verbal feedback from the trainer), visual (delivered as demonstrations of correct motion by the trainer), and haptic (delivered as guiding nudges by the trainer to ensure the subject reaches the desired range of motion).

The first design of the athletic demonstrator platform uses primarily postural information to deliver score-based feedback as it is often the most important category of feedback for performance in sports motion, but other categories of feedback will be added to the platform to allow for a richer set of information on performance with the potential to improve motor learning.

3.3.2. *Case study on quantification of feedback*

Once the categories of feedback and modalities of feedback in motor learning were determined, the next step was to determine how an automated system can observe and provide feedback on an individual's performance in each of these categories. In the athletic demonstrator platform, the system has access to a 3-dimensional representation of a user's motion as a time-series dataset extracted from fused Kinect and IMU data. In a similar project, "Autonomous Training Assistant" [56], we found that all three categories of performance can be inferred from this data by comparing to expert motion. To determine when to provide feedback, it is useful to set a threshold at which an error can be identified in each category. In other words, once the user's motion deviates from the expert's motion by a targeted amount, feedback can be given to correct that motion. We call this method "tolerance thresholding", and it can be used to refine our definition of each modality of feedback in the following ways:

Postural data may be described as the way in which an individual's joint angles, and for the relevant joints in a motion, relate to one another and to an expert's joints in 3D space. At any given point in time, the Kinect can determine the location and angle of a user's shoulders, elbows, wrists, knees, and other joints for coarse postural adjustment (fine postural adjustment requires more sensitive recording mechanisms which may be implemented, for example, as wearable sensors). For each joint related to the posture of a motion, we can define postural performance as the proximity of a user's joint angle to that of the expert at that point in time, adjusted using Dynamic Time Warping (DTW) methods to ensure the two are equally scaled. The tolerance threshold for posture can then be defined as the maximum amount, in degrees, that a subject's joint is allowed to deviate from the expert's joint for the motion to be considered "correct". Deviation beyond this point can be considered an error and feedback can be given accordingly.

For progression, a system can observe the trajectory of a user's motion and compare it to the expert's trajectory for assessment, noting the proximity of the two in time-adjusted 3D space. It would be difficult to perform this assessment for every recorded data point of a motion in real-time; instead, only the most essential points, i.e., "critical points", representing the shape and form of the motion may be observed. An arc, for example, can be represented as a progression of five points in space. At these points along the motion, a user's data point can be compared to an expert's data point using a standard 3-dimensional distance measure. The tolerance threshold can be defined for progression as the maximum allowed distance between two critical points for a motion to be "correct" at that point in time.

Finally, for pacing, a system can observe the rate at which a user progresses from the start to the end of a motion, compared to an expert. The difference between these two forms the user's error in pacing. A user's motion is allowed to be slower or faster than the expert's motion up to a specific tolerance threshold to be considered "correct". Beyond this range, feedback is necessary. Note that in this case a system must specify, as the trainer does in our first example above, whether the motion is slower or faster than the desired rate so that the user can make adjustments in the proper direction.

The final design of the athletic demonstrator platform can use the above metrics, determined through the quantification of the trainer's feedback in the case study, to form a detailed profile of a user's performance for a motion.

3.3.3. *Case study for individual preference*

To determine the effects of individual preference on the effectiveness of a modality in a multimodal feedback scenario in motor learning, a study was designed with the case study subject wherein a multimodal environment with the Autonomous Training Assistant was presented. In this environment, the subject was asked to complete a series of simple motor exercises with two feedback conditions. In the first condition, modalities (haptic, audio, visual) were assigned to feedback categories (posture, progression, pacing) based on the mapping suggested by the review of Sigrist *et al.* [53] for concurrent multimodal feedback. In the second condition, the subject was able to choose the mapping based on individual preference. The subject then completed a series of three basic martial arts exercises (umbrella motion, twirl motion, and witik motion) assigned by the subject's martial arts trainer using the Autonomous Training Assistant interface for each condition.

Each exercise was completed in a 2-minute interval with breaks in-between, and a longer break between the two conditions to prevent fatigue and minimize learning effects. The subject's performance was measured in each category using error rate in

May 25, 2017 6:11:52pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

each of the three performance categories. It was found that in the preference condition, the subject performed significantly better in categories that were mapped differently by preference, while performance in unchanged categories of feedback remained the same between conditions. This improvement held consistently across all three exercises, suggesting that individual preference in multimodal feedback selection may have an effect on performance in multimodal training environments. Furthermore, it was observed that, in both conditions, the subject would focus on a single modality of feedback over the other modalities in the presence of multimodal feedback. In this case, the subject seemed to focus on haptic feedback as indicated by an increased responsiveness to feedback in modality.

Further studies on a larger scale using the athletic demonstrator platform can help determine whether these observations are generalizable across a variety of users and motor training exercises. Currently, the platform is capable of providing terminal feedback using a score system to indicate performance on a motor exercise via expert data, which is purely visual. Haptic feedback will be added to the platform through the introduction of wrist-worn vibrotactile motors to guide the user at regular intervals through movements as initially described in [55]. Furthermore, rhythmic audio cues will be added to accompany both the demonstration and attempt screens of the platform to help the user compare the rhythm of their motion to that of an expert as an additional form of evaluation.

4. Wait Time and Queue Estimation

The objective of this project is to enrich the fan experience by providing access to wait times at restrooms and concession stands via a mobile app. Such a technology will allow fans to maximize their time watching and enjoying a game rather than waiting in long lines during the course of a game. We adopt a computer vision based approach to count the number of people in a queue. We assume the presence of cameras in strategic locations in the vicinity of restrooms and concession stands; the video feed from these cameras is analyzed to accurately estimate the count of people in the queues. Once the count is obtained, wait times can be obtained from the average service time per person.

Counting the number of objects in an image is a problem of paramount practical importance. It arises in myriads of real-world applications including crowd behavior monitoring, security and surveillance, medical imaging and developing infrastructures for smart cities, among others. Counting is often posed as a supervised learning problem, where a regression function is learned directly from some global image features to the number of objects in it. The regression-based algorithms depict commendable performance in counting the number of objects in images. However, they necessitate a large amount of manually annotated data from human oracles to train the regression models. This is an expensive process in terms of time, labor and human expertise. Further, annotating an image for object counting requires much more time and effort than annotating an image for a face recognition or an object

Fig. 6. Two images with ground truth object counts.

recognition application, for instance. Figure 6 shows two images of pedestrians in a shopping mall and in an outdoor walkway, together with the corresponding ground truth counts. It is evident that hand-labeling such images with counts of objects is an extremely tedious task and highly prone to annotation errors. Thus, while annotating a face/object image requires only a cursory glance, counting objects is much more laborious and demands significantly more time, effort and concentration from a human oracle. It is therefore a significant challenge to obtain a large amount of labeled training images with the exact counts of the number of objects in them. In this paper, we propose a novel learning framework, with the following two features, to address this fundamental problem: (i) the first feature, binary user feedback, relaxes the requirement of exact count of objects as labels; (ii) the second feature, active sampling, aims to reduce the amount of labeled training data (and hence, the amount of manual effort) required to induce a regression model. These are detailed below:

4.1. *Binary user feedback*

We present a general learning framework which requires only binary (yes/no) feedback from the user. During each instance of interaction, the human user is presented with an image and a threshold (an integer) and he merely has to say whether the number of objects in the image is greater than the threshold or not. Providing such an input is extremely easy; it is also less prone to human errors as the number of objects in an image needs to be compared only against a given threshold every time.

In order to quantitatively compare the two types of user feedback: exact (where the exact count of the number of objects needs to be provided) and binary (where only a *yes/no* response needs to be provided about whether the count of objects is greater than a given threshold), we conducted experiments on 15 users. Each user was shown a sequence of four random images, one from each of the following

May 25, 2017 6:11:53pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

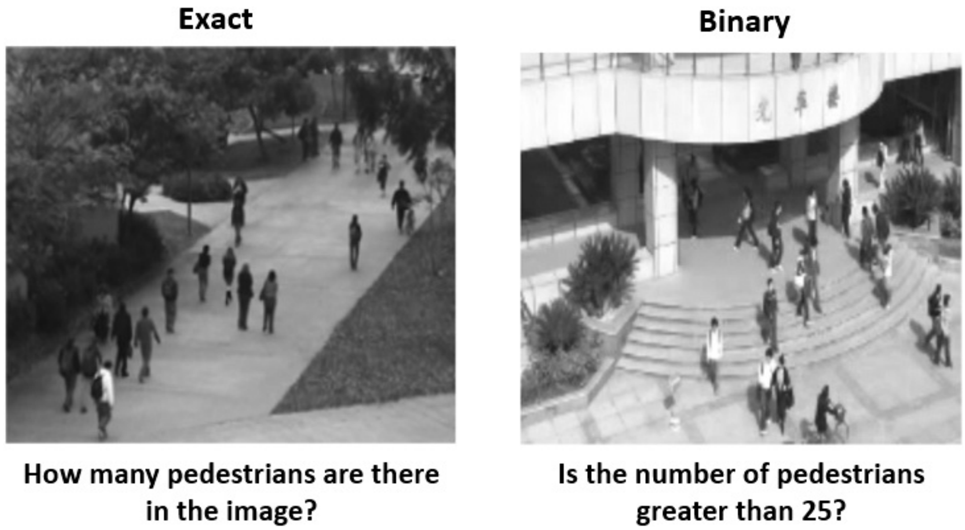


Fig. 7. Exact and binary annotation examples (the thresholds for the binary annotations in the experiment were computed using our algorithm).

datasets: the Mall [37], the UCSD pedestrians [10], the Fudan [57] and the TRANCOS [42]. These datasets contain images captured under challenging real-world conditions. For the first two images, the user was asked to provide exact annotations and for the next two, binary annotations (the thresholds for the binary annotations in the experiment were computed using our algorithm and is detailed in Sec. 4.4). Sample images are shown in Fig. 7.

We computed the response time (time taken to annotate an image) for both exact and binary annotations; we also requested each user to provide an overall score between 1 (extremely difficult) and 10 (extremely easy) about the ease of binary annotation over exact annotation. The results are depicted in Table 6. We note that the binary feedback requires much lesser user interaction time than the exact annotations. Moreover, as evident from the scores, users were much more comfortable with the binary annotations since it does not involve the strenuous task of counting the exact number of objects in an image. In summary, binary feedback provides an extremely appealing user interaction model for the vision based object counting application.

Table 6. User study results on exact and binary annotations.

Annotation type	Mean response time (seconds)	Mean score
Exact	20.98 ± 4.34	9.26 ± 0.88
Binary	11.32 ± 4.74	

4.2. *Active sampling*

Active learning algorithms have gained popularity in reducing human annotation effort for training machine learning models. When exposed to large amounts of unlabeled data, such algorithms automatically identify the salient and prototypical instances which can augment maximal information to the underlying models [50]. While serial-query based active learning algorithms query a single unlabeled sample at a time, batch mode active learning (BMAL) techniques query a batch of samples simultaneously for manual annotation and are effective in utilizing the presence of multiple labeling oracles. BMAL has been successfully used in a variety of computer vision applications such as face and facial expression recognition [8], image and video retrieval [31] and image clustering [22] among others. In this work, we exploit batch sampling algorithms to identify the exemplar images that need to be queried for labels, from vast amounts of unlabeled image samples. This can tremendously reduce the human annotation effort required to induce the regression learner, as only the exemplar samples identified by the algorithm need to be labeled manually. To the best of our knowledge, this is the first research effort to address the problem of active data selection with binary user feedback in the context of vision-based object counting. Although validated on object counting in this paper, the proposed algorithm is generic and can be used in any regression-based application where the exemplar instances need to be selected from large amounts of unlabeled data and a model needs to be trained based on binary user feedback.

4.3. *Related work*

In this section, we present a survey of vision based object counting methodologies as well as a brief survey of active learning.

Vision-based Object Counting: Unsupervised learning techniques have been used to address the vision-based object counting problem. They mostly rely on grouping objects based on self-similarities [1] or motion similarities [44]. However, these techniques are limited in their counting accuracy, which has paved the way for supervised learning approaches for counting. Detection-based supervised algorithms attempt to train object detectors (e.g. pedestrian detectors) to localize the individual object instances within an image; the count is then estimated as the number of localized objects. Common approaches of detection-based counting include non-maximum suppression [16], generative techniques [5] and blob tracking [26] among others. Fusion based approaches have also been explored for people counting [32] which rely on multiple sources of information (low confidence head detections, repetition of texture elements and frequency domain analysis) to estimate counts of individuals in extremely crowded images. However, all these techniques need to solve object detection, which is a challenging computer vision problem, especially for overlapping and occluded instances.

The regression-based counting techniques avoid solving the hard detection problem and attempt to learn a mapping directly from some global image feature to the number of objects in it. Cho *et al.* [13] used edge features together with background subtraction and reported promising performance while estimating crowd density using a neural network. Kong *et al.* [34] performed feature normalization in a neural network model to deal with perspective projection and camera orientation and proposed a viewpoint invariant approach to count pedestrians. Chen *et al.* [12] recently proposed a scalable multi-output regression model to estimate people count in spatially localized regions. Marana *et al.* [38] postulated that images of low density crowds tend to present coarse textures while images of dense crowds present fine textures; a self-organizing neural network was used to extract features from such images for crowd density estimation. Lempitsky and Zisserman [35] proposed to recover a density function F as a real function of the pixels in an image I, so that integrating F over the entire image yields the count of the number of objects in it. Very recently, deep learning algorithms have been exploited to count the number of objects in an image [49].

Active Learning: Active learning is a well-studied problem in machine learning. Several techniques have been developed over the last several years and a review of these can be found in [50]. In a typical pool-based batch mode active learning (BMAL) setting, the learner is exposed to a pool of unlabeled instances and it iteratively queries batches of samples for annotation. Initial BMAL techniques were largely based on heuristic measures such as maximizing the diversity of the selected samples, computed as their distance from the decision hyperplane [6]. More recently, optimization based strategies have been proposed which have been shown to outperform the heuristic approaches. Hoi *et al.* [30] used the Fisher information matrix as a measure of model uncertainty and proposed to query the set of points that maximally reduced the Fisher information. Semi-supervised BMAL algorithms have also been explored in the context of SVMs, where a kernel function was first learned from a mixture of labeled and unlabeled samples, which was then used to identify the informative and diverse examples through a min-max framework [31]. Guo and Schuurmans [25] proposed a discriminative strategy that selected a batch of points which maximized the log-likelihoods of the selected points with respect to their optimistically assigned class labels and minimized the entropy of the unselected points in the unlabeled pool. Guo also proposed a batch mode active learning scheme which maximized the mutual information between the labeled and unlabeled sets and was independent of the classification model [24]. Chakraborty *et al.* [9] introduced an active matrix completion algorithm to select the most informative queries to complete a low rank matrix. Researchers have also explored theoretical properties of active learning and have established concrete mathematical bounds on the expected number of queries to achieve a given error rate [15].

While active learning has been extensively studied in a variety of computer vision applications, it has been comparatively much less explored for object counting. Loy *et al.* [37] proposed a regression based active learning algorithm (m-landmark) for crowd counting, which was based on computing the normalized Graph Laplacian

L followed by k-means clustering. However, the algorithm did not consider binary user feedback about the object count. The Elastic Net algorithm proposed by Tan *et al.* [57] was based on a similar clustering strategy; however, it was more focused on selecting a promising set of initial training samples for semi-supervised learning, rather than active learning. In this paper, we propose a novel object counting algorithm which can identify the exemplar unlabeled samples for manual annotation and requires only binary feedback from human oracles. We now describe the proposed framework.

4.4. *Proposed framework*

Let $\{x_{l1}, x_{l2}, \ldots, x_{lN}\}$ be the set of N instances, which are labeled with their exact counts $Y = \{y_1, y_2, \ldots, y_N\}$ and let $\{x_{u1}, x_{u2}, \ldots, x_{uM}\}$ be the set of the unlabeled instances. Our objective is to select a batch containing k most informative unlabeled samples from the unlabeled set, obtain their binary annotations from the human oracle and use that to predict the labels of all the unlabeled samples. This task can be decomposed into the following two research questions (**RQs**): (i) How can we use active learning to select the k most informative samples from the unlabeled set for binary user annotation? and (ii) Given the current labeled set containing the exact counts and the set of k newly selected samples from the unlabeled set with binary (*yes/no*) annotations, how can we predict the labels of all the unlabeled samples?

Conventional regression-based counting algorithms (such as the ridge regression or the support vector regression) require the exact count of the number of objects in each data sample and are hence unsuitable for our application. Given our problem set-up, we need a framework which can incorporate inequality constraints (greater or less than a given threshold) in estimating the count of objects in images. An alternative strategy is to pose regression learning as the problem of completing a low rank matrix [9]. Further, Marecek *et al.* [39] recently proposed a matrix completion algorithm under interval uncertainty, to impute the missing entries of a data matrix in the presence of equality and inequality constraints. In this paper, we exploit matrix completion algorithms for the problem of object counting from binary user feedback.

4.4.1. *Matrix completion*

The data collected in most computer vision/machine learning applications are structured in the form of matrices. For instance, in a classification/regression problem, each row represents a data sample, with corresponding label(s) and each column denotes a feature; in a recommendation system, the data is represented in the form of a matrix, where each row is a user, each column is an object and the corresponding entry represents the rating given by the particular user to that object. Due to flaws in the feature acquisition process or the unwillingness of subjects to disclose personal information, the collected data often contains missing entries, which can bias results, reduce generalizability and lead to erroneous conclusions.

Matrix completion algorithms attempt to reconstruct a matrix from a set of partially observed entries and are of immense practical importance [7, 45]. Such techniques have also been exploited to address classification and regression problems [23]. The fundamental assumption is that the stacked matrix $Z = [Y^0; X^0]$ containing the label matrix Y^0 and the feature matrix X^0 is jointly low rank. The missing entries in the matrix correspond to the labels of the unlabeled samples and are estimated using matrix completion algorithms. It is posed as the following optimization:

$$\begin{array}{ll} \min\limits_{Z} & \mathrm{rank}(Z) \\ s.t.\colon Z_{ij} = E_{ij}, & \forall i, j \in E \end{array} \tag{2}$$

where E is the set of the observed entries. Several methods have been devised to efficiently optimize this problem. The Fixed Point Continuation (FPC) method in particular, is an iterative algorithm consisting of a gradient step and a shrinkage step in each iteration with guaranteed monotonic convergence [23].

4.4.2. *RQ1: Active sampling of the unlabeled data instances*

Our object counting framework is based on the theory of matrix completion, necessitating an active learning framework within the matrix completion paradigm. Chakraborty *et al.* [9] recently proposed the *Active Matrix Completion* algorithm to identify the missing entries in a partially observed matrix, which are the most informative to reconstruct the original matrix. The fundamental idea was to compute a measure of uncertainty of prediction of every missing entry in the incomplete data matrix; the top uncertain entries were then queried for manual annotation. Three strategies were presented to quantify the prediction uncertainty of each missing entry in the incomplete matrix: (i) *Conditional Gaussians*, which assumes that the set of missing entries conditioned on the set of observed entries follows a multivariate normal distribution; the mean and covariance matrix of the conditional distribution are computed from the given data and the diagonal elements of the covariance matrix quantifies the variance (uncertainty) associated with each imputation; (ii) *Query by Committee (QBC)*, which uses a committee of matrix completion algorithms to impute the missing entries and quantifies the prediction uncertainty of a particular entry as the level of disagreement among the committee; and (iii) *Committee Stability*, which is similar to QBC and quantifies the prediction uncertainty using the regularity of predictions of a particular entry from an ensemble of predictors.

In this work, we used the QBC algorithm for active instance sampling due to its promising performance in matrix completion [9] and active learning in general, its strong theoretical properties [51] and ease of implementation. Specifically, a committee of matrix completion algorithms were applied on the partially observed data matrix to impute the missing values. The variance of prediction (among the committee members) of each missing entry was taken as a measure of uncertainty of that entry. The top k uncertain entries were then queried for manual annotation. We used the following three commonly used matrix completion algorithms as members of our committee:

***k*-NN:** The k-nearest neighbor algorithm identifies the k most similar features to the current one with a missing value and uses the average of these k nearest neighbors as an estimate for the missing entry [28].

EM: This method imputes the missing values using the Expectation Maximization (EM) algorithm [28]. An iteration of the EM algorithm involves two steps. In the E step, the mean and covariance matrix are estimated from the data matrix (with the missing entries filled with zeros or estimates from the previous M step); in the M step, the missing value of each data column is imputed with their conditional expectation values based on the available entries and the estimated mean and covariance. The mean and the covariance are re-estimated based on the newly completed matrix and the process is iterated until convergence.

SVD: Singular value decomposition (SVD) is a standard method for matrix completion based on low-rank approximation [28]. In this method, initial guesses are first provided to the missing data values. SVD is then applied to obtain a low rank approximation of the filled-in data matrix. The missing values are then updated based on their corresponding values in the low rank estimation. SVD is applied to the updated matrix again and the process is iterated until convergence.

4.4.3. *RQ2: Counting with binary user feedback*

In our framework, the user provides only binary (*yes/no*) annotations to the unlabeled samples selected using active learning. This necessitates a matrix completion scheme that can handle inequality constraints apart from equality constraints, as in Eq. (2). The *MACO* algorithm proposed by Marecek *et al.* [39] uses alternating parallel co-ordinate descent to complete a matrix in the presence of equality, lower bound and upper bound constraints. Let X be the $m \times n$ matrix to be reconstructed. Suppose that for the elements $(i, j) \in E$, we have equality constraints, for the elements $(i, j) \in L$ we have lower bounds and for the elements $(i, j) \in U$, we have upper bounds. Completing the matrix can thus be posed as the following optimization:

$$\begin{aligned} &\min_{X \in \Re^{m \times n}} && \operatorname{rank}(X) \\ &s.t.\text{: } X_{ij} = X_{ij}^{E}, && \forall (i,j) \in E \\ &X_{ij} \geq X_{ij}^{L}, && \forall (i,j) \in L \\ &X_{ij} \leq X_{ij}^{U}, && \forall (i,j) \in U \end{aligned} \tag{3}$$

The problem in Eq. (3) is NP-hard, even with $U = L = \phi$ [39]. A popular heuristic enforces low rank in a synthetic way by writing X as a product of two matrices, $X = AB$, where $A \in \Re^{m \times r}$ and $B \in \Re^{r \times n}$ Hence, X is of rank at most r. The alternating parallel co-ordinate descent algorithm to solve the above optimization is outlined in Algorithm 1 (please refer to [39] for more detailed derivations).

Algorithm 1. The MACO algorithm for matrix completion under equality and inequality constraints

Require: E, L, U, X^E, X^L, X^U, rank r
1: choose $A \in \Re^{m\times r}$ and $B \in \Re^{r\times n}$
2: **for** $k = 1, 2, \ldots,$ **do**
3: choose a random subset $\hat{S}_{row} \in \{1, 2, \ldots, m\}$
4: **for** $i \in \hat{S}_{row}$ **do**
5: choose $\hat{r} \in \{1, 2, \ldots, r\}$ uniformly at random
6: update $A_{i\hat{r}} = A_{i\hat{r}} + \delta_{i\hat{r}}$, where $\delta_{i\hat{r}}$ is computed using co-ordinate descent
7: **end for**
8: choose a random subset $\hat{S}_{column} \in \{1, 2, \ldots, n\}$
9: **for** $j \in \hat{S}_{column}$ **do**
10: choose $\hat{r} \in \{1, 2, \ldots, r\}$ uniformly at random
11: update $B_{\hat{r}j} = B_{\hat{r}j} + \delta_{\hat{r}j}$, where $\delta_{\hat{r}j}$ is computed using co-ordinate descent
12: **end for**
13: **end for**
14: **return** $m \times n$ matrix AB

In our object counting application, the initial training set containing the exact counts of the objects forms the set E. The MACO algorithm is used only with the set E to derive estimates of the missing labels of the unlabeled samples. These estimates are used as thresholds for binary user query; the binary user feedback on the selected unlabeled samples constitute the sets L and U. The MACO algorithm is used again with the sets E, L and U to estimate the missing labels of the unlabeled samples. The pseudo-code of our algorithm is presented in Algorithm 2.

4.5. *Experiments and results*

Datasets and Feature Extraction: We used four challenging datasets from different application domains to study the performance of the proposed framework: (i) the Mall dataset [37] containing video frames collected using a publicly accessible webcam for crowd counting and profiling research; (ii) the UCSD Pedestrian dataset [10], which contains videos of pedestrians on UCSD walkways, taken from a stationary camera; (iii) the Fudan Pedestrian dataset [57], which contains video frames captured at one side entrance of Guanghua Tower, Fudan University, Shanghai, China; and (iv) the TRaffic ANd COngestionS (TRANCOS) dataset [42], a novel benchmark for (extremely overlapping) vehicle counting in traffic congestion situations. All these datasets are captured under challenging real-world conditions with severe inter-object occlusions, varying crowd densities from sparse to crowded, as well as diverse activity patterns (static and moving crowds) under varying illumination conditions at different times of the day. Sample images from these datasets

Algorithm 2. The proposed active object counting algorithm with binary user feedback

Require: The labeled initial training set $\{x_{l1}, x_{l2}, \ldots, x_{lN}\}$, their ground truth counts $Y = \{y_1, y_2, \ldots, y_N\}$, the unlabeled set $\{x_{u1}, x_{u2}, \ldots, x_{uM}\}$, batch size k, rank parameter r

1: Form the stacked matrix $Z = [Y; X]$; the labels of the unlabeled samples constitute the missing entries
2: Form the equality set E from the given label set Y
3: Apply the MACO algorithm (Algorithm 1) using the constraint set E to derive estimates of the missing labels of the unlabeled samples
4: Use the QBC algorithm [9] on Z to select the k most informative unlabeled samples
5: Query the binary labels of the selected k samples with respect to the thresholds given by their label estimates computed in Step 3
6: Form the lower bound and upper bound constraint sets L and U from the binary user feedback
7: Apply the MACO algorithm again using the constraint sets E, L and U to complete the missing entries of the matrix
8: **return** The labels of the unlabeled samples

are shown in Figs. 6 and 7. The histogram of oriented gradients (HOG) feature [14] was used as the descriptor of each image frame due to its established performance in computer vision tasks.

4.5.1. *Experimental setup*

Each dataset was randomly divided into a labeled training set and an unlabeled set. The batch size was set at 10% of the dataset size (as detailed in Table 7). A batch of samples was queried from the unlabeled set, appended to the labeled set and the performance was evaluated on the complete unlabeled set. We studied the performance of binary annotations (where the user merely provides *yes/no* answers as to whether the number of objects in an image is greater than a given threshold) for both random selection as well as active sampling of the unlabeled samples. In random sampling, a batch of samples was selected at random from the unlabeled set

Table 7. Dataset details.

Dataset	Number of samples	Batch size
Mall	2000	200
UCSD	2000	200
Fudan	1500	150
TRANCOS	1200	120

for annotation while in active sampling, the proposed active learning framework was used to select the unlabeled samples for annotation. We also studied the performance of exact annotations (where the user provides the exact count of the number of objects in an image) for both random and active sampling. For exact annotations, we used only the equality constraint set E in the MACO algorithm; the lower and upper bound constraint sets L and U were empty since the user provided the exact counts. We used the mean-squared error (MSE) on the unlabeled set as the evaluation metric in our work. For each dataset, we studied the performance with different sizes of the initial training set, from 10% to 50% in steps of 10%.

4.5.2. *Regression using matrix completion*

Matrix completion algorithms have been used successfully to address regression problems [9]. We first studied the performance of matrix completion for the regression-based object counting problem. We used three common regression algorithms — ridge regression, kernelized ridge regression and support vector regression — as comparison baselines. The results on the four datasets are reported in Table 8 (in each experiment, 70% of the data was used for training and 30% for testing).

Thus, matrix completion provides comparable performance to other counting techniques. However, our method has the flexibility of incorporating binary user feedback in contrast to other methods which need the exact counts for model training.

4.5.3. *Active counting with binary feedback*

The results for the four datasets are reported in Tables 9–12. All the results were averaged over 5 runs (with different labeled and unlabeled sets) to rule out the effects of randomness. **BaseMSE** denotes the mean squared error using the current training data (before sample selection and annotation); **Binary Ann** denotes the MSE corresponding to binary annotations while **Exact Ann** denotes the MSE corresponding to exact annotations. **Random** denotes the case when the unlabeled samples are selected at random for user annotation while **Active** denotes the case when active sampling is used to select the unlabeled samples.

Table 8. Comparison of matrix completion (MC) against regression algorithms. Error metric: Mean squared error.

Dataset	MC	Ridge regression	Kernel ridge regression	Support vector regression
Fudan	2.09	1.51	3.0	1.54
Mall	9.26	4.60	7.20	4.69
TRANCOS	115.99	86.30	147.08	89.84
UCSD	30.01	6.54	7.13	6.78

May 25, 2017 6:12:01pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

Table 9. MSE comparison results on the Mall dataset. Lower values denote better performance.

		Binary Ann		Exact Ann	
Train %	BaseMSE	Random	Active	Random	Active
10	918.14	629.59	505.75	548.27	394.24
20	813.29	475.91	317.65	422.42	225.02
30	714.46	370.50	209.1	344.80	148.78
40	612.29	334.80	154.68	322.69	111.5
50	510.05	253.69	90.99	264.71	58.35

Table 10. MSE comparison results on the UCSD dataset. Lower values denote better performance.

		Binary Ann		Exact Ann	
Train %	BaseMSE	Random	Active	Random	Active
10	780.71	527.55	526.32	361.89	347.33
20	704.85	407.78	315.24	269.77	144.10
30	611.38	311.61	247.05	218.77	136.78
40	525.84	245.6	176.75	158.16	70.09
50	443.26	218.43	151.79	152.72	73.79

Table 11. MSE comparison results on the Fudan dataset. Lower values denote better performance.

		Binary Ann		Exact Ann	
Train %	BaseMSE	Random	Active	Random	Active
10	36.21	30.79	28.43	28.51	24
20	33.43	27.92	19.81	22.41	15.84
30	31.89	19.11	12.61	15.85	10.99
40	28.72	17.64	9.8	14.62	8.21
50	24.23	16.67	7.86	14.93	7.04

Table 12. MSE comparison results on the TRANCOS dataset. Lower values denote better performance.

		Binary Ann		Exact Ann	
Train %	BaseMSE	Random	Active	Random	Active
10	1.3 e+03	984.14	911.94	861.94	829.24
20	1.23 e+03	685.56	640.86	574.59	515.72
30	1.1 e+03	548.57	483.86	440.63	371.47
40	952.78	394.25	344.82	313.92	271.21
50	794.23	296	266.1	227.15	197.07

May 25, 2017 6:12:01pm WSPC/214-IJSC 00275 ISSN: 1793-351X Page Proof

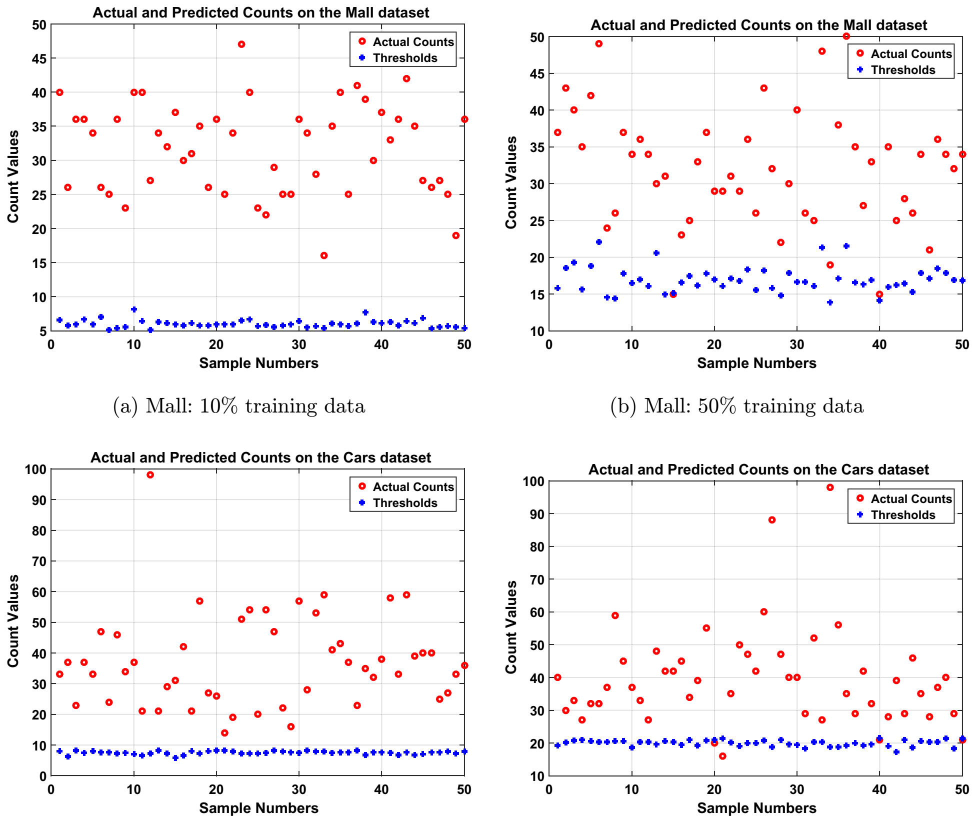
We first note that the MSE reduces with increasing size of the initial training set, which is intuitive. We also note that the algorithm based on binary annotations delivers much better performance compared to the baseline error. This corroborates the usefulness of the proposed framework in tremendously reducing the error rate by exploiting only binary feedback from the human user. Moreover, active sampling successfully identifies the salient and exemplar unlabeled instances and further improves the error rate over random selection in a binary user feedback setting. The same pattern is evident for different sizes of the initial training set and for all the datasets, depicting the generalizibility of our framework. Thus, while conventional learning frameworks can operate only with data annotated with the exact counts of objects, our framework offers more flexibility and ease of interaction between the user and the machine. From these results, we conclude that the proposed framework can be immensely useful to boost the accuracy of an object counting system while minimizing the labeling burden on human oracles.

The algorithm based on exact annotations produces better performance compared to that based on binary annotations. This is intuitive, as exact annotation provides more information to the underlying machine learning models. As before, active instance sampling further reduces the error rate compared to random sampling. More importantly, we note that active sampling with binary user annotations often provides comparable results (and sometimes, even outperforms) random sampling with exact annotations, which is the conventional method to address the counting problem. This depicts the merit of our algorithm in tremendously reducing human annotation effort with minimal effect on the counting accuracy.

4.5.4. *Threshold study*

In our framework, a threshold is first computed by the algorithm and the user provides a binary feedback as to whether the number of objects in the image is greater or less than the threshold (the threshold is computed as the current label estimate of the sample in question). Thus, the user annotation time depends on the threshold computed by the MACO algorithm. If the threshold is close to the actual count of objects, the annotation time will be higher and vice versa. In this experiment, we study the thresholds computed by our algorithm on 50 random unlabeled samples for 10% and 50% initial labeled training data. The results on the Mall and TRANCOS datasets are depicted in Fig. 8.

We note that with 10% labeled training data, the thresholds computed are coarse and thus, the binary annotation time will be low. As the percentage of training data increases, the prediction accuracy increases and consequently, the computed thresholds are much closer to the actual counts. Hence, the binary annotation time will be almost similar to the absolute annotation time, since an exhaustive count of all the objects will be necessary for accurate annotations. Our framework is therefore most useful in the initial stages of learning, when the amount of labeled training data



(a) Mall: 10% training data

(b) Mall: 50% training data

(c) TRANCOS: 10% training data

(d) TRANCOS: 50% training data

Fig. 8. Study of the threshold computed in our algorithm. Best viewed in color.

is scarce. With abundant labeled data, absolute annotations are more advantageous. A hybrid framework can be envisioned, where binary annotations are used in the initial stages and absolute annotations in the later stages of learning. This will be investigated as part of future research.

5. Discussion and Future Work

Three fan enrichment projects in the scope of the Smart Stadium for Smarter Living initiative were presented. These projects targeted improved safety (Crowd Understanding), fan engagement (Athletic Demonstrator Platform), and efficiency/convenience (Wait Time/Queue Estimation). Through use of smart stadia as testbeds, the manageable size and heterogeneity of these testbeds enabled practical trials while still providing a useful environment to explore challenges of scalability and realtimeness. Preliminary results presented here demonstrate the potential of these technologies for smart city solutions.

As part of future work, we are developing and deploying new projects for the smart stadia. These projects include smart solutions to address issues of congestion and difficulty parking during large events at the stadia; game-within-a-game halftime interactions to enrich fan engagement; and projects that target important priority areas of energy efficiency and sustainability. We are also investigating the use of the Athletic Demonstrator Platform as a low-cost, accurate platform to augment traditional athlete training intended for use outside of sessions involving the trainer or coach. Moreover, future studies with the platform are being planned to observe how this feedback can be integrated over time to adapt to a user's proficiency level, and how this integration can differ between individuals and various types of movement. One such study will investigate how multimodal feedback delivery may be tuned for fast sports motion interaction as opposed to slower rehabilitative movements, and how the type of movement may be inferred from the nature of the expert data samples.

Acknowledgments

The authors thank Intel Corporation, National Science Foundation, Arizona State University, and Dublin City University for their funding support. This material is partially based on work supported by: Intel Corporation under grant Joint Path Finding (JPF) Proposal: Smart Stadium and Smart Living Research; and National Science Foundation under Grant No. 1069125.

References

[1] N. Ahuja and S. Todorovic, Extracting texels in 2.1d natural textures, in *IEEE International Conference on Computer Vision*, 2007.

[2] A. Ahmadi, E. Mitchell, F. Destelle, M. Gowing, N. E. O'Connor, C. Richter and K. Moran, Automatic activity classification and movement assessment during a sports training session using wearable inertial sensors, in *Proc. 11th International Conference on Wearable and Implantable Body Sensor Networks*, 2014, pp. 98–103.

[3] A. Ahmadi, F. Destelle, D. Monaghan, K. Moran, N. E. O'Connor, L. Unzueta and M. T. Linaza, Human gait monitoring using body-worn inertial sensors and kinematic modelling, in *Proc. IEEE SENSORS*, 2015, pp. 1–4.

[4] A. E. Atwater, Biomechanics of overarm throwing movements and of throwing injuries, *Exercise and Sport Sciences Reviews* **7**(1) (1979) 43–86.

[5] O. Barinova, V. Lempitsky and P. Kohli, On the detection of multiple object instances using hough transforms, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010

[6] K. Brinker, Incorporating diversity in active learning with support vector machines, in *International Conference on Machine Learning*, 2003.

[7] E. Candes and T. Tao, The power of convex relaxation: Near-optimal matrix completion, in *IEEE Transactions on Information Theory* **56**(5) (2010) 2053–2080.

[8] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan and J. Ye, Active batch selection via convex relaxations with guaranteed solution bounds, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(10) (2015) 1945–1958.

[9] S. Chakraborty, J. Zhou, V. Balasubramanian, S. Panchanathan, I. Davidson and J. Ye, Active matrix completion, in *IEEE International Conference on Data Mining*, 2013.

[10] A. Chan, Z. Liang and N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[11] D. K. Chan, C. Lonsdale, P. Y. Ho, P. S. Yung and K. M. Chan, Patient motivation and adherence to postsurgery rehabilitation exercise recommendations: The influence of physiotherapists autonomy-supportive behaviors, *Arch Phys Med Rehabil* **90**(12) (2009) 1977–1982.

[12] K. Chen, C. Loy, S. Gong and T. Xiang, Feature mining for localized crowd counting, in *British Machine Vision Conference*, 2012.

[13] S. Cho, T. Chow and C. Leung, A neural-based crowd estimation by hybrid global learning algorithm, in *IEEE Transactions on Systems, Man and Cybernetics* **29**(4) (1999) 535–541.

[14] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[15] S. Dasgupta, Coarse sample complexity bounds for active learning, in *Advances of Neural Information Processing Systems*, 2005.

[16] C. Desai, D. Ramanan and C. Fowlkes, Discriminative models for multi-class object layout, in *IEEE International Conference on Computer Vision*, 2009.

[17] F. Destelle, A. Ahmadi, N. E. O'Connor, K. Moran, A. Chatzitofis, D. Zarpalas and P. Daras, Low-cost accurate skeleton tracking based on fusion of Kinect and wearable inertial sensors, in *Proc. 22nd European Signal Processing Conference*, 2014, pp. 371–375.

[18] F. Destelle, A. Ahmadi, K. Moran, N. E. O'Connor, N. Zioulis, A. Chatzitofis, D. Zarpalas, P. Daras, L. Unzueta, J. Goenetxea and M. Rodriguez, A multi-modal 3D capturing platform for learning and preservation of traditional sports and games, in *Proc. 23rd ACM International Conference on Multimedia*, 2015, pp. 747–748.

[19] J. E. Deutsch, M. Borbely, J. Filler, K. Huhn and P. Guarrera-Bowlby, Use of a low-Cost, commercially available gaming console (Wii) for rehabilitation of an adolescent with cerebral palsy, *Phys. Ther.* **88**(10) (2008) 1196–1207.

[20] S. Essid, D. Alexiadis, R. Tournemenne, M. Gowing, P. Kelly, D. Monaghan, P. Daras, A. Drmeau and N. E. O'Connor, An advanced virtual dance performance evaluator, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 2269–2272.

[21] E. Farella, L. Benini, B. Ricc and A. Acquaviva, MOCA: A low-power, low-cost motion capture system Based on integrated accelerometers, *Adv. MultiMedia* **2007**(1) (2007) 11.

[22] C. Fu and Y. Yang, A batch-mode active learning SVM method based on semi-supervised clustering, in *Intelligent Data Analysis*, 2015.

[23] A. Goldberg, X. Zhu, B. Recht, J. Xu and R. Nowak, Transduction with matrix completion: Three birds with one stone, in *Advances of Neural Information Processing Systems*, 2010.

[24] Y. Guo, Active instance sampling via matrix partition, in *Advances of Neural Information Processing Systems*, 2010.

[25] Y. Guo and D. Schuurmans, Discriminative batch mode active learning, in *Advances of Neural Information Processing Systems*, 2007.

[26] I. Haritaoglu, D. Harwood and L. Davis, W4: real-time sureveillance of people and their activities, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(8) (2000) 809–830.

[27] T. Hassner, Y. Itcher and O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior, in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, p. 16.

[28] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown and D. Botstein, Imputing missing data for gene expression arrays, in Technical Report, Stanford University, 1999.

[29] S. Hodgkinson, Is your city smart enough? Digitally enabled cities and societies will enhance economic, social, and environmental sustainability in the urban century, in *OVUM Report*, 2011.

[30] S. Hoi, R. Jin, J. Zhu and M. Lyu, Batch mode active learning and its application to medical image classification, in *International Conference on Machine Learning*, 2006.

[31] S. Hoi, R. Jin, J. Zhu and M. Lyu, Semi-supervised SVM batch mode active learning for image retrieval, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[32] H. Idrees, I. Saleemi, C. Seibert and M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[33] P. K. Tra, K. Pong and R. Bowden, An improved adaptive background mixture model for real-time tracking with shadow detection, in *Video-Based Surveillance Systems*, eds. P. Remagnino, G. A. Jones, N. Paragios and C. S. Regazzoni (Springer, 2002), pp. 135–144.

[34] D. Kong, D. Gray and H. Tao, Counting pedestrians in crowds using viewpoint invariant training, in *British Machine Vision Conference*, 2005.

[35] V. Lempitsky and A. Zisserman, Learning to count objects in images, in *Neural Information Processing Systems*, 2010.

[36] H. Liu, X. Wei, J. Chai, I. Ha and T. Rhee, Realtime human motion control with a small number of inertial sensors, in *Proc. Symposium on Interactive 3D Graphics and Games*, 2011, pp. 133–140.

[37] C. Loy, S. Gong and T. Xiang, From semi-supervised to transfer counting of crowds, in *IEEE International Conference on Computer Vision*, 2013.

[38] A. Marana, S. Velastin, L. Costa and R. Lotufo, Estimation of crowd density using image processing, in *Image Processing for Security Applications*, 1997.

[39] J. Marecek, P. Richtarik and M. Takac, Matrix completion under interval uncertainty, in *European Journal of Operational Research* **256**(1) (2017) 35–43.

[40] M. Marsden, K. McGuinness, S. Little and N. E. OConnor, Holistic features for real-time crowd behaviour anomaly detection, in *Proc. IEEE International Conference on Image Processing*, 2016, pp. 918–922.

[41] U. Nations, World urbanization prospects: The 2014 revision, highlights. Department of Economic and Social Affairs, in *Population Division, United Nations*, 2014.

[42] R. Olmedo, B. Jimnez, R. Sastre, S. Bascn and D. Rubio, Extremely overlapping vehicle counting, in *Iberian Conference on Pattern Recognition and Image Analysis*, 2015.

[43] N. Otsu, A threshold selection method from gray-level histograms, *Automatica* **11** (285296) (1975) 2327.

[44] V. Rabaud and S. Belongie, Counting crowded moving objects, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[45] B. Recht, A simpler approach to matrix completion, in *Journal of Machine Learning Research* **12** (2011) 3413–3430.

[46] D. J. Reinkensmeyer and S. J. Housman, "If I cant do it once, why do it a hundred times?": Connecting volition to movement success in a virtual environment motivates people to exercise the arm after stroke, in *Proc. Virtual Rehabilitation 2007*, 2007, pp. 44–48.

[47] D. Roetenberg, H. Luinge and P. Slycke, Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors, in *Xsens Motion Technologies BV*, Technical Report, 2009.
[48] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics* **6**(2) (1978) 461–464.
[49] S. Segui, O. Pujol and J. Vitria, Learning to count with deep object features, in *IEEE Conference Computer Vision and Pattern Recognition Workshop*, 2015.
[50] B. Settles, Active learning literature survey, in Technical Report 1648, University of Wisconsin-Madison, 2010.
[51] H. Seung, M. Opper and H. Sompolinsky, Query by committee, in *Workshop on Computational Learning Theory*, 1992.
[52] J. Shao, K. Kang, C. C. Loy and X. Wang, Deeply learned attributes for crowded scene understanding, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4657–4666.
[53] R. Sigrist, G. Rauter, R. Riener and P. Wolf, Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review, *Psychon Bull. Rev.* **20**(1) (2013) 21–53.
[54] N. Skjret, A. Nawaz, T. Morat, D. Schoene, J. L. Helbostad and B. Vereijken, Exercise and rehabilitation delivered through exergames in older adults: An integrative review of technologies, safety and efficacy, *International Journal of Medical Informatics* **85**(1) (2016) 1–16.
[55] R. Tadayon, T. McDaniel, M. Goldberg, P. M. Robles-Franco, J. Zia, M. Laff, M. Geng and S. Panchanathan, Interactive motor learning with the autonomous training assistant: A case study, in *Human-Computer Interaction: Interaction Technologies* (Springer International Publishing, 2015), pp. 495–506.
[56] R. Tadayon, T. McDaniel and S. Panchanathan, Autonomous training assistant: A system and framework for guided at-home motor learning, in *Proc. 18th International ACM SIGACCESS Conference on Computers and Accessibility*, 2016, pp. 293–294.
[57] B. Tan, J. Zhang and L. Wang, Semi-supervised elastic net for pedestrian counting, in *Pattern Recognition* **44**(10–11) (2011) 2297–2304.
[58] C. Tomasi and T. Kanade, Detection and tracking of point features, in Technical Report CMU-CS-91-132, 1991.
[59] Smart cities council — Transportation [Online]. Available: http://readinessguide.smartcitiescouncil.com/readiness-guide/transportation-0.
[60] G. Welch and E. Foxlin, Motion tracking: No silver bullet, but a respectable arsenal, *IEEE Computer Graphics and Applications* **22**(6) (2002) 24–38.
[61] M. Windolf, N. Gtzen and M. Morlock, Systematic accuracy and precision analysis of video motion capturing systems exemplified on the Vicon-460 system, *Journal of Biomechanics* **41**(12) (2008) 2776–2780.
[62] Z. Zhang, Microsoft Kinect sensor and its effect, *IEEE MultiMedia* **19**(2) (2012) 4–10.
[63] L. Zhang, J. C. Hsieh, T. T. Ting, Y. C. Huang, Y. C. Ho and L. K. Ku, A Kinect based golf swing score and grade system using GMM and SVM, in *Proc. 5th International Congress on Image and Signal Processing*, 2012, pp. 711–715.