Autonomous computational-intelligence based behaviour recognition in security and surveillance

Louis G. Clift[a], Jason Lepley[b], Hani Hagras[a], and Adrian F. Clark[a]

[a]School of Computer Science & Electronic Engineering, University of Essex, Colchester, UK
[b]Leonardo MW Ltd, Sigma House, Basildon, Essex, UK

ABSTRACT

This paper presents a novel approach to sensing both suspicious, and task-specific behaviours through the use of advanced computational intelligence techniques. Locating suspicious activity in surveillance camera networks is an intensive task due to the volume of information and large numbers of camera sources to monitor. This results in countless hours of video data being streamed to disk without being screened by a human operator. To address this need, there are emerging video analytics solutions that have introduced new metrics such as people counting and route monitoring, alongside more traditional alerts such as motion detection. There are however few solutions that are sufficiently robust to reduce the need for human operators in these environments, and new approaches are needed to address the uncertainty in identifying and classifying human behaviours, autonomously, from a video stream. In this work we present an approach to address the autonomous identification of human behaviours derived from human pose analysis. Behavioural recognition is a significant challenge due to the complex subtleties that often make up an action; the large overlap in cues results in high levels of classification uncertainty. False alarms are significant impairments to autonomous detection and alerting systems, and over reporting can lead to systems being muted, disabled, or decommissioned. We present results on a Computational-Intelligence based Behaviour Recognition (CIBR) that utilises artificial intelligence to learn, optimise, and classify human activity. We achieve this through extraction of skeleton recognition of human forms within an image. A type-2 Fuzzy logic classifier then converts the human skeletal forms into a set of base atomic poses (standing, walking, etc.), after which a Markov-chain model is used to order a pose sequence. Through this method we are able to identify, with good accuracy, several classes of human behaviour that correlate with known suspicious, or anomalous, behaviours.

Keywords: Behaviour Recognition, Activity Recognition, Human Activity Recognition, Automated Surveillance, Computer Vision, Computational Intelligence, Machine Learning, Fuzzy Logic

1. INTRODUCTION

The number of CCTV systems in use today world-wide is immense, providing around-the-clock monitoring of places such as buildings and open spaces. Existing Commercial Off The Shelf (COTS) surveillance systems offer useful autonomous video analytical features such as motion detection, allowing them to detect, highlight, track and record movement in the field of view. Event triggers can be configured to raise alerts via email, SMS (text message) or set off specific physical alarm systems based on pre-defined thresholds. High-end camera manufacturers have integrated these features directly within the camera,[1] moving the intelligent sensing to the 'edge'. Edge-processing[2] is a concept aimed at decentralising heavy processing tasks to move the computation near to the source of the data. With cloud-based remote services requiring all of the data to be transmitted across the internet to remote services hosted by external companies, a series of concerns arise such as security and data transmission fees. Such edge-processing is attractive in video surveillance, principally because of bandwidth constraints, privacy concerns and reducing the demand on central processing.

With a seemingly ever-increasing demand for CCTV systems, countless hours of video are being recorded straight to disk without ever being reviewed by a human operator, or simply not recorded. Novel CCTV solutions

Further author information:
A: E-mail: lclift@essex.ac.uk
B: E-mail: Jason.Lepley@leonardocompany.com

are seeking to reduce the burden of human CCTV operators through the use of autonomous monitoring systems which seek to raise alerts in the event of an incident, and analysis systems based around computer vision have much to offer here. Newer systems extend conventional capabilities with ones such as Automatic Number Plate Recognition (ANPR),[3] object detection and classification (*e.g.*, identifying buses, cars, cyclists and people),[4] face detection and pedestrian-[5] or vehicle-counting.[6] However, detecting specific behaviours remains a complex and difficult problem due to the vast range of behaviours and minor subtleties between a normal motion and a suspicious behaviour. Behaviour analysis is a challenging task which can encompass anything from sitting down to attempting to steal a car. What is it that separates a person unlocking their vehicle with a traditional key versus a criminal attempting to gain entry? Real-world behaviours are difficult to detect, even for a trained eye.

Advances in sensor technology have introduced accessible depth sensors to form camera systems known as RGB-D or depth cameras. The Microsoft Kinect is one of the leading sensors in this area: its first generation combined an infra-red projector and camera with a visual-band camera and its second iteration is based around a time-of-flight system. Although originally designed for interactive gaming, the Kinect has been adopted by researchers due to its ability to provide feature-rich 3D data,[7–10] including an accurate human skeleton model which enables researchers to develop novel solutions to traditionally difficult tasks. In previous work,[11] a second-generation Microsoft Kinect was used to provide a skeleton to a custom-designed computational intelligence system to classify the poses and simple behaviours of the people within the field of view.

Although RGB-D sensors provide exceptionally high resolution at relatively short ranges (0.5–4.5 m), introducing RGB-D industrial surveillance applications would be impractical, not only due to the size of the sensor package but the cost of such a system would become cost-prohibitive. Conventional CCTV camera enclosures require a more compact sensor package than can be achieved with RGB-D sensors. In this work we explore the steps required to remove the reliance on RGB-D sensing whilst being able to continue to experiment with an Interval Type-2 Fuzzy Logic (IT2FL) classifier-based[11] on skeletal data.

The remainder of this contribution is structured as follows. Section 2 defines our distinction of poses, motions, behaviours and finally activities or actions. We have deliberately avoided taking a single approach to the problem of behaviour recognition; instead the aim has been to split the task into key phases and design components to tackle each of the challenges on the route to behaviour identification. Section 4 details the classification system used to identify atomic poses. Section 5 discusses the early experimental results at the time of writing. Section 6 explains the next stages in moving from pose to motion before section 7 draws the paper to its conclusions.

2. DEFINING POSES VERSUS BEHAVIOURS

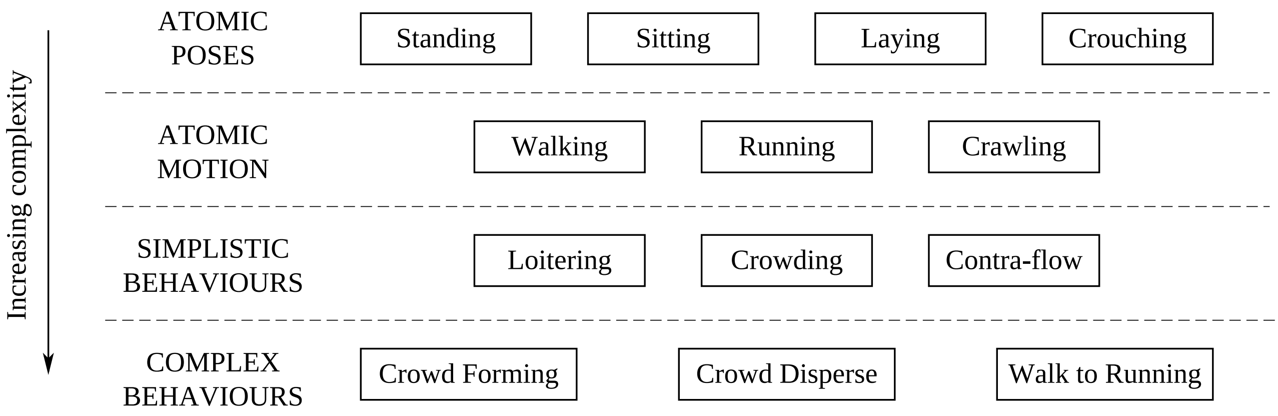

Figure 1. A selection of common behaviours which can be visually identified by a trained CCTV operator. These behaviours can be categorised into poses, motions and behaviours.

Human activity/action recognition (HAR) and pattern recognition aims to identify the actions or objectives of either an individual or collection of people (crowds).[12] Many areas of Computer Vision, Data Mining and smart device tracking[13] can be used to learn the behaviours (or habits) of a particular person. An 'activity' is a very abstract term which can be used to describe a sequence of actions which form a complex event. Before moving on to the technical contribution of this work we first breakdown activities into distinct behaviours.

Figure 1 begins with the most basic of behaviours, e.g. standing, sitting. We define these behaviours as a pose, a posture that can be observed at any given time point within a time span (video/image sequence). Given that any of the more complex behaviours can include a collection of these poses, we define them as atomic poses, the most fundamental of positions in which the subject is at rest, not moving. However, complex behaviours typically involve movement, therefore we introduce the category of atomic motion, activities which do not require a context to explain the behaviour, such as walking and running. The two are kept separate as each band of category has a non-linear increase in technical complexity. Security specific activities such as loitering or crowd assessments don't emerge until the later stages as behaviours. Moving from the atomic activities (pose/motion) to behaviours requires combinations of the atomic sets as well as different classification techniques.

3. SYSTEM OVERVIEW

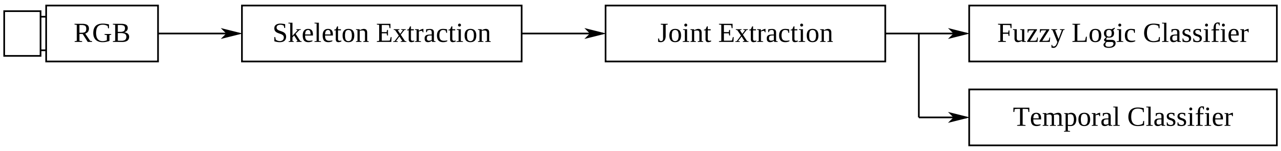

Figure 2. Block-level diagram of classification system

The Computational-Intelligence based Behaviour Recognition (CIBR) system is comprised of several key stages. The first stage is the visual input processor. A single high-definition (720p or above) camera source is used in either live or pre-recorded mode to supply frames to the CIBR system. This can be either a directly connected USB camera or via a network or UNIX socket (in the case of capture-card driven sources). Each frame is encoded into a matrix and passed on to the next stage for extraction.

Extracting skeletons (or humans) from images is a significant research area within the fields of computer vision and signal processing, as there are a myriad of challenges coupled with detecting a human such as background segmentation, identification and spine extraction. Although there are many algorithms with seek to solve this challenge, a recent solution, OpenPose[14–16] has successfully and robustly enabled this work to remove the requirement for a RGB-D sensing. The next block is first of the main work of this research. The raw joint data produced by the skeleton extractor are processed into angular information before passing into the fourth stage of the work-flow.

The classifiers form the core of the CIBR system. There are a number of classifiers in use, each with a specific role within the system. The Type-2 Fuzzy Logic classification system is by far the biggest which is responsible for identifying the atomic poses in every frame. This information is shared with the other classifiers to establish a more complex understanding of the behaviour of the subjects seen in the video.

The final stage (not shown in figure 2) is the data presentation and storage element. Specific behaviours need to logged and others require evidence such as the frames for human review. The aim is construct a tool that aids CCTV operators, not automate their work-flow. A human is significantly more capable in determining context and intent than that of any automated system.

4. TYPE-2 FUZZY LOGIC POSE CLASSIFICATION

With the aid of OpenPose, skeletons are extracted from live camera sources. The output from the library is in raw pixel coordinates for each joint, which are converted to angular measurements for input to the fuzzy system. Angular data enables the classifier to become scale invariant and recognise poses regardless of distance to the camera. The fuzzy classification system determines the pose of each person visible to OpenPose. There are a number of limitations with using OpenPose which are discussed later in the paper.

Although ones gait is said to be as unique as a fingerprint, the same cannot be said about poses. Just standing still can present a number of poses, yet all require the same linguistic label as one is still ‘standing’. Fuzzy Logic excels in this area and was designed to handle uncertain data in complex real-world applications.

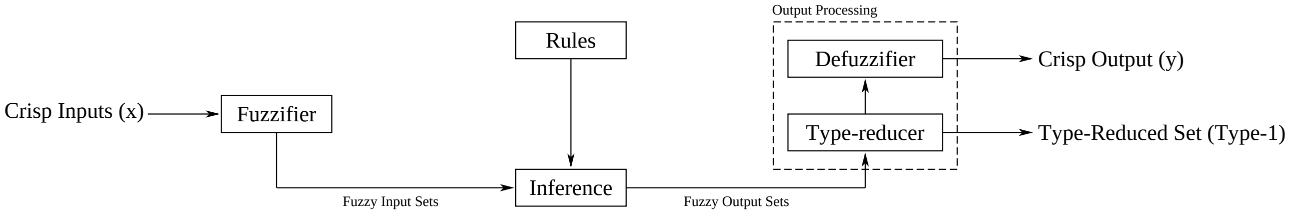

Figure 3. Typical type-2 fuzzy system

Figure 3 demonstrates a top-level overview of a type-2 fuzzy logic system. Real world data inputs are known as crisp inputs. In the case of this work these represent the angular joints from a 10-point model. Typically, crisp values are raw readings from an input or sensor and contain a large amount of uncertainty and noise. Each input is fuzzified using an input fuzzy set. The input set will depend on the type of fuzzy system being deployed. A fuzzy set consists of a series of membership functions which encode the probability that a given input falls into a specific linguistic label. Linguistic labels enable engineers to group a series of values into a specific state. For example if an input was reading distance between a robot and a wall, values between 0.5 m – 1.0 m could be said to be at a medium distance with 1.0 m – 1.5 m as far. Fuzzy input sets remove the hard edge boundary at 1.0 m. Readings between 0.8 m – 1.2 m would assessed by the input set and passed to the inference stage for classification.

Type-1 Fuzzy systems can handle limited uncertainty but can be still fairly sensitive to noisy boundary cases. Type-2 systems have uncertainty designed directly into the membership functions through the use of Interval Type-2 Fuzzy Sets (IT2FS) as seen in figure 4. Each membership, here assigned an arbitrary label, has both an upper and lower membership function (shape) assigned to each label. This enables a more reliable account for variation in input signals.

Once crisp input has been fuzzified, it produces a series of labels and their associated firing strengths. A rulebase (shown as ‘Rules’ in figure 3) is a set labels and a desired set of output labels. The inference engine uses binary operations to assess the fit for the incoming fuzzy state and the expected outputs. A rule base is ‘learnt’ during a training stage whereby known inputs are fed to the system so that the rules can be setup by the fuzzy system in a supervised learning approach.

Once passed through inference stage, the rule is then passed through the output fuzzy set to obtain the final output (in fuzzy form). A type-2 system uses type reduction to convert the upper and lower outputs into a reduced set of values, which would be expected from a standard type-1 fuzzy logic system. As with a type-1 system, this output is passed through a defuzzifier to obtain a crisp output which a machine can use to provide control, feedback or decision making.

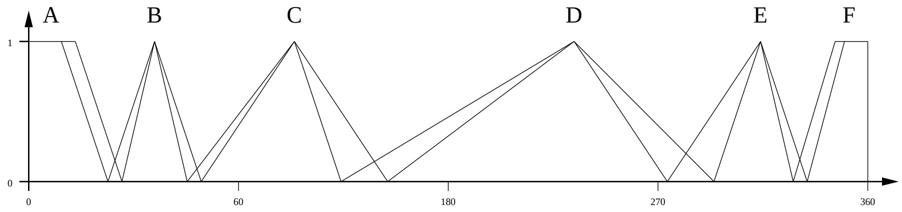

Figure 4. Type-2 membership function definition for an individual input joint angle

There are three core skeleton models to select within OpenPose offering 15, 18 or 25 body key-points each with varying accuracy and performance trade-offs. Regardless of model mode selected, in this work we reduce this down to just 10 joints from the skeleton model produced from OpenPose. These include the wrist, elbows, shoulders, hips and knees. Each joint is passed through its own fuzzy set which in turn can have its individual membership functions tuned. The input fuzzy set seen in figure 4 is the starting set for the live fuzzy system before optimisation.

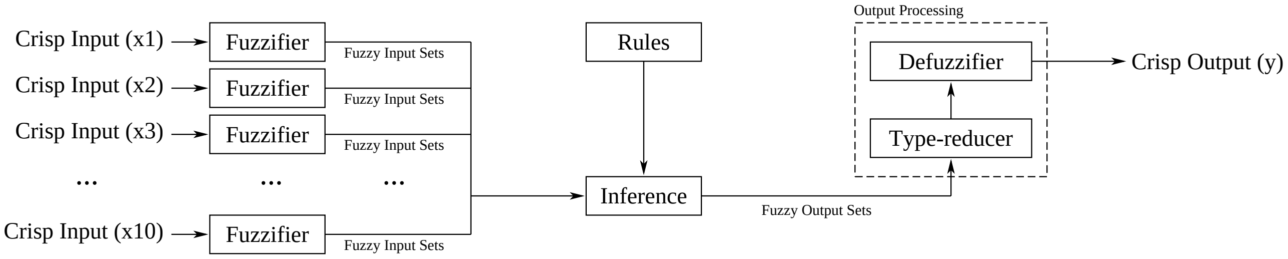

Figure 5. Modified type-2 fuzzy system

Each input fuzzy set then produces a linguistic label for the supplied joint angle. Figure 5 shows a visual representation of the expaned input system compared with the traditional fuzzy workflow. Each label produced from the input sets are combined together to form an antecedent pattern, akin to a sentence along with a singular firing strength, which is evaluated as one by the inference engine and rule base. For example if a particular pose was to output a label of 'A' for the left wrist, it be entered into the pattern along with the rest of the angles, e.g. 'A,A,B,E,A,A,C,B,B,D'. Each rule is encoded in the same format. This approach enables string comparison instead of binary operations within inference (rule matching).

The CIBR system is comprised of multiple fuzzy systems, one for each of the desired atomic poses. Each system receives a copy of the joint data. The output from each system is a score of the likelihood that the incoming set of angles matches that specific pose. A higher-level function selects the fuzzy system producing the highest score as the classified pose.

5. EXPERIMENTAL RESULTS

We have explored a number of different problems with this system, encompassing both conventional surveillance ones and other pose-oriented ones such as semaphore gesture identification. Table 1 summarises results from three experiments, and we present both the initial results obtained while developing the system and those from formal K-fold cross-validation exercises.

Table 1. Effect of K-Fold cross-validation for supervised learning

	Number of frames		**Development fold**		**10-fold cross-validation**	
Dataset	Training	Testing	Training	Testing	Training	Testing
sit, stand	148	84	100.0%	92.31%	100.0%	94.87%
sit, stand, walk	585	148	98.89%	65.00%	86.00%	71.67%
Semaphore 5	35	100	100.0%	68.00%	95.00%	82.86%

The first experiment is a simple binary classification of poses into the classes `sit` and `stand`. The fuzzy system was able to classify its training data perfectly and obtained $> 92\%$ accuracy. A formal experiment using 10-fold cross validation (four for training, one for testing) again yielded perfect results on the training fold and almost 95% mean accuracy on unseen test imagery.

The second experiment extended the first one with an additional class, `walk`. Initial results suggested that training was near-perfect but the performance on unseen test data was poor. However, a formal test using 10-fold cross-validation produced somewhat better mean performance on unseen test data.

Our final experiment was to classify five semaphore poses. Our initial experiment yielded perfect performance on the training data and moderate performance on unseen test data, while a formal 10-fold cross-validation experiment produced somewhat better mean performance on unseen test imagery, $\sim 83\%$.

We regard these performance figures as promising. It is expected that classifications that occur in a single video frame in a sequence of consistent classifications can simply be discarded as being incorrect as a human subject cannot change their pose for a single frame of a 25 Hz video stream. Hence, we anticipate performance on complete videos to be somewhat better than the results reported in table 1.

Experiments into the limitations and operational constraints of the pose analysis revealed that the strength of the system is under-pinned by the skeleton extraction library. During testing it was observed that certain poses were undetectable by OpenPose. One such pose is a person laying down. The pose can, under certain conditions, be detected by rotating the image prior to analysis. Although a solution to the problem, it has dramatic effect on performance as every frame will require multiple passes through skeleton extraction. These findings are based on the library as of May-June 2018. Further testing and experimentation is required since the introduction of a new extraction model within the library.

6. MOVING FROM POSE TO MOTION

The classifiers presented thus far have performed well at single pose identification, however do not directly extend to the next category, namely motion detection. In limited trials these motion sets were passed into the IT2FS fuzzy system, although the classifier was able to classify these movements it dramatically reduces the accuracy and reliability of the atomic pose classification, introducing large errors. An alternate approach to extending the classifier is to instead use the classified output across multiple frames. Walking could be described as a sequence of ‘standing’ poses across n-frames. Through this assumption that atomic motion can essentially be ‘paused’ to reveal an atomic pose, we can use time-based classifiers such as decision-trees or more complex Markov-chain modelling to label these new behaviours.

Moving from pose to motion introduces a new challenge to the system. Detecting singular poses can run independently for each frame. To understand motion, each frame in a sequence needs to be tracked and mapped back on to a path such that poses in a sequence all belong to the same individual. Microsoft are able to track up to 6 people within its field of view utilising the Kinect version 2. OpenPose is capable of supplying large numbers of skeletons leaving the tracking to the application layer. In this work motion is tracked between frames by calculating the centre of mass of the skeleton and then analysing the motion of this point between images. The centre of mass is calculated through finding the centre between the shoulders and hips, as these joints form the core. Limbs are excluded as these joints could influence the central point even if the subject is remaining stationary.

Target tracking is a well established problem. Tracking a single person as they move through the field of view is challenging, but solvable with the above approach. Tracking a target amongst other skeletons becomes significantly more complex. In order to move from classifying atomic sets to behaviours requires the ability to track and follow the same individual for duration of the time in the field of view.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a task-oriented approach to analysing seemingly simple poses and motions which form complex human behaviours. Early system tests in a controlled environment demonstrate a high level of accuracy is achievable using a type-2 fuzzy logic system from a 2D image and arteficial intelligence developed skeleton models. Further work is required to refine the pose sets and to identify the limitations of the solution through a range of real-world pose sets, target tracking scenarios alongside more complex environments involving multiple people crossing through the scene.

A 2D-based behavioural analysis system is more suited to rapid deployment and system upgrades without needing specialist cameras to be fitted to the surveillance environment, reducing both cost and installation complexity. 3D RGB-D systems would require specialised re-engineering before deployment. Our work has the added advantage of being deployable at either a centralised data centre or at the edge, through the use of dedicated edge-processing units.

ACKNOWLEDGMENTS

This work is supported through a Knowledge Transfer Programme (KTP) between the University of Essex and Leonardo MW Ltd. Knowledge Transfer Partnerships are a part government-funded programme to encourage collaboration between businesses and universities in the United Kingdom.

REFERENCES

[1] Sony, "Sony intelligence video analytics - distributed enhanced processing architecture (depa)," *Sony Techinical Documentation* **1**, 14 (dec 2006).

[2] Garcia Lopez, P., Montresor, A., Epema, D., Datta, A., Higashino, T., Iamnitchi, A., Barcellos, M., Felber, P., and Riviere, E., "Edge-centric computing: Vision and challenges," *SIGCOMM Comput. Commun. Rev.* **45**, 37–42 (Sept. 2015).

[3] Patel, C., Shah, D., and Patel, A., "Automatic number plate recognition system (anpr): A survey," *International Journal of Computer Applications* **69**(9), 21–33 (2013).

[4] Chen, Z., Ellis, T., and Velastin, S. A., "Vehicle detection, tracking and classification in urban traffic," in [*Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*], 951–956, IEEE (2012).

[5] Kim, G.-J., Eom, K.-Y., Kim, M.-H., Jung, J.-Y., and Ahn, T.-K., "Automated measurement of crowd density based on edge detection and optical flow," in [*Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on*], **2**, 553–556, IEEE (2010).

[6] Sina, I., Wibisono, A., Nurhadiyatna, A., Hardjono, B., Jatmiko, W., and Mursanto, P., "Vehicle counting and speed measurement using headlight detection," in [*Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on*], 149–154, IEEE (2013).

[7] Liu, A.-A., Nie, W.-Z., Su, Y.-T., Ma, L., Hao, T., and Yang, Z.-X., "Coupled hidden conditional random fields for rgb-d human action recognition," *Signal Processing* **112**, 74–82 (2015).

[8] Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D., "Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments," in [*In the 12th International Symposium on Experimental Robotics (ISER*], Citeseer (2010).

[9] Xia, L., Chen, C.-C., and Aggarwal, J. K., "Human detection using depth information by kinect," in [*Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*], 15–22, IEEE (2011).

[10] Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D., "Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments," *The International Journal of Robotics Research* **31**(5), 647–663 (2012).

[11] Yao, B., Lepley, J. J., Peall, R., Butler, M., and Hagras, H., "Recognition of complex human behaviours using 3d imaging for intelligent surveillance applications," in [*Emerging Imaging and Sensing Technologies*], **9992**, 99920H, International Society for Optics and Photonics (2016).

[12] Kim, E., Helal, S., and Cook, D., "Human activity recognition and pattern discovery," *IEEE Pervasive Computing/IEEE Computer Society [and] IEEE Communications Society* **9**(1), 48 (2010).

[13] Lara, O. D., Labrador, M. A., et al., "A survey on human activity recognition using wearable sensors.," *IEEE Communications Surveys and Tutorials* **15**(3), 1192–1209 (2013).

[14] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y., "Realtime multi-person 2d pose estimation using part affinity fields," in [*CVPR*], (2017).

[15] Simon, T., Joo, H., Matthews, I., and Sheikh, Y., "Hand keypoint detection in single images using multiview bootstrapping," in [*CVPR*], (2017).

[16] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y., "Convolutional pose machines," in [*CVPR*], (2016).