Understanding Movement and Interaction: an Ontology for Kinect-based 3D Depth Sensors

Natalia Díaz Rodríguez[1], Robin Wikström[2], Johan Lilius[1], Manuel Pegalajar Cuéllar[3], and Miguel Delgado Calvo Flores[3]

[1] Turku Centre for Computer Science (TUCS), Department of IT, Åbo Akademi University, Turku, Finland
[2] IAMSR, Åbo Akademi University, Turku, Finland
{ndiaz, rowikstr, jolilius}@abo.fi
[3] Department of Computer Science and Artificial Intelligence,
University of Granada, Spain.
{manupc, mdelgado}@decsai.ugr.es

Abstract. Microsoft Kinect has attracted great attention from research communities, resulting in numerous interaction and entertainment applications. However, to the best of our knowledge, there does not exist an ontology for 3D depth sensors. Including automated semantic reasoning in these settings would open the doors for new research, making possible not only to track but also understand what the user is doing. We took a first step towards this new paradigm and developed a 3D depth sensor ontology, modelling different features regarding user movement and object interaction. We believe in the potential of integrating semantics into computer vision. As 3D depth sensors and ontology-based applications improve further, the ontology could be used, for instance, for activity recognition, together with semantic maps for supporting visually impaired people or in assistance technologies, such as remote rehabilitation.

Keywords: Ontology, Kinect, Human Activity Modelling and Recognition, Ubiquitous Computing

1 Introduction

Recently there has been a spark in developments in the field of smart spaces and ubiquitous computing, especially regarding applications using affordable sensors. One of these sensors is the Microsoft Kinect device, originally intended as an add-on for the Xbox 360 video console, which enables user interaction through movements and voice, instead of using a controller. However, the sensor attracted a lot of interest from the R&D communities, as Kinect can be reprogrammed for other purposes than purely entertainment.

The main goal with ubiquitous spaces is to work towards an ideal environment where humans and surrounding devices interact effortlessly [1]. For this to be realized, context-awareness is key. Semantic technologies have shown to be successful, among other areas, in context representation and reasoning, which can serve in object tracking and scene interpretation [2] and in human activity recognition [3]. We believe that semantic modelling of human movement and interaction could greatly benefit existing data-driven (e.g., computer vision) approaches, increasing context-awareness and potentially, activity recognition rates.

One of the most challenging areas within *UbiComp* is Activity Recognition. Using vision based techniques has substantial disadvantages, as most of them store the images, and become intrusive and privacy compromising. Since 3D depth sensors do not store the image itself, but a skeleton structure, they add an advantage towards traditional data-driven approaches [4] (HMM, SVM, etc.)

To the best of our knowledge, there does not exist any automated semantic reasoning for modelling movement and interaction within computer vision technologies and 3D depth sensors (e.g. Kinect). The rest of the paper is structured as follows. Section 2 presents related work in computer vision and semantic approaches, Section 3 describes our ontology proposal for modelling body movement, and Section 4 exemplifies its usage. Section 5 concludes and gives some future research directions.

2 Related Work

Due to Kinect multimodal features such as gesture and spoken commands, different UbiComp applications have been recently developed. For instance, the combination of Kinect with an airborne robot [5] to enable automatic 3D modelling and mapping of indoor environments.

An interesting initiative in this area is Kinect@Home[4][6], a crowd-sourcing project for large 3D datasets of real environments to help robotics and computer vision researchers, through vast amounts of images, to improve their algorithms. Another project, *Kinect Fusion* [7], allows for real-time 3D reconstruction and interaction using point-based 3D depth sensor data. An application example is touch input enabled arbitrary surfaces.

In the Semantic Web, ontologies represent the main technology for creating interoperability at a semantic level. This is achieved by creating a formal illustration of the data, making it possible to share and reuse the ontology all over the Web. Ontologies formulate and model relationships between concepts in a given domain [8]. The following example illustrates with OWL 2 axioms the activity $TakeMedication$, that can serve to monitor an elder: $NataliaTakingMedication \equiv isPerformedBy.(Natalia \sqcap performsAction$ $(OpenPillCupboard \sqcap (TakeObject \sqcap actionAppliesTo\, some\, NataliasMedication) \sqcap$ $(TakeObject \sqcap actionAppliesTo\, some\, Glass) \sqcap FillGlassWithWater \sqcap Drink))$.

In [9] ontology-based annotation of images and semantic maps are realized within a framework for semantic spatial information processing. An XML description language for describing the physical realization of behaviours (speech

[4] Kinect@Home http://www.kinectathome.com/

and behaviour) is the Behavior Markup Language (BML) [5], which allows representation of postures and gestures for controlling verbal and nonverbal behavior of (humanoid) embodied conversational agents (ECAs). However, to the best of our knowledge, there is no current solution integrating the performance power of computer vision technologies, together with a formal semantic representation of the user, its movement and interaction with the environment, to achieve automatic knowledge reasoning. In next section we propose an ontology for combining data-driven and knowledge-based paradigms.

3 An Ontology for modelling movement and interaction with 3D depth sensors

We propose an ontology to distinguish among human movement, human-object interaction and human-computer interaction. The Kinect ontology[6] aims at representing 3D depth sensor information generally, but at this stage it is based upon two main Kinect modules. The first and most basic one is Kinect Core, and represents the Natural User Interface (NUI), which is the core of the Kinect for Windows API, and represents the most relevant concepts from Kinect Interaction and Kinect Fusion APIs [10]. The second module of the ontology consists of practical extensions for modelling and recognizing human activity.

3.1 Kinect Core Ontology, Kinect Interaction and Kinect Fusion

The *Kinect Sensor* class represents the camera device, its current location, orientation and frames. A Kinect Sensor associates a *3D Model* with the user' skeleton.

A Kinect 3D *Volume* is characterized through its size and voxel resolution.

Kinect Audio supports a microphone mode, beamforming and source localization (which can be identified through a direction or language). A *Speech* is recognized by a *Speech Recognition Engine*. The latter allows creation of customized grammars for recognition of user commands with a confidence threshold parameter for each grammar.

Kinect Interaction provides several ways to interact with a Kinect-enabled application. The natural gestures, as a way of touch-free user interactions, allow the sensor to operate in a range of 0.4 to 3-4 m. The types of interaction are modelled with gestures (gripping, releasing, pushing and scrolling) (Fig. 1). This class generates interaction streams which are bound to a control, i.e., an action that allows computer interaction. A *Control* is an action performed when an interaction gesture is recognized. The set of interactive controls are classified on video, images or text. An *Interaction Stream* represents the supply of interaction frames as they are generated by a *KinectInteraction*. Each *InteractionFrame* has a timestamp.

[5] BML: `http://www.mindmakers.org/projects/bml-1-0/wiki`

[6] Kinect Ontology: `http://users.abo.fi/rowikstr/KinectOntology/`

Kinect distinguishes among two types of *Tracking Modes*, *default* or *seated*. Both modes can track 2 out of 6 users, but only one can be active at once.

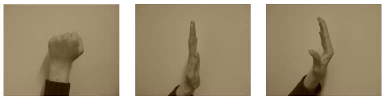

Fig. 1. Some available interaction gestures: a) Grip b) Release c) Press

By using ontology-based modelling, a *SeatingUser* can be defined as:

$SeatingUser \equiv User \sqcap isTracked \sqcap (SeatedTrackingMode\, isActive)$.
$StandingUser \equiv User \sqcap isTracked \sqcap (DefaultTrackingMode\, isActive)$.
$TrackedUser \equiv User \sqcap isTracked \sqcap ((DefaultTrackingMode\, or\, SeatedTrack-ingMode) isActive)$.
$InteractingUser \equiv User \sqcap isTracked \sqcap (hasArm\, some\, Arm) \sqcap (hasHand\, some\, Hand) \sqcap (hasInteractionMode\, some\, (GrippingInteractionMode\, or\, ReleasingInter-actionMode\, or\, PressingInteractionMode))$.

Kinect' *Skeleton* class identifies a *User* and is represented with a bone and joint hierarchy, which refers to the ordering of the bones defined by the surrounding joints. Our ontology allows to express relations concerning bones and joints, where the bone rotation is stored in a bone's child joint, e.g., the rotation of the left hip bone is stored in the *HipLeft* joint (See Fig. 2-right)[7]. The skeletal tracking includes rotations of each bone joint and orientations of each bone.

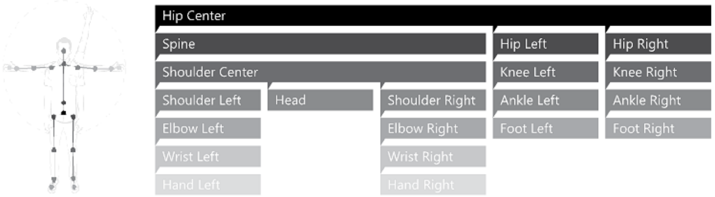

Fig. 2. Left) Skeleton, bones and joints. Right) Joints hierarchy. [10]

The *Hand* class has a set of properties that represent its state, e.g., the user the hand belongs to, whether the hand is primary for that user, whether the hand is interactive, gripping or pressing. *Arms*, in the same way, are provided with an arm state.

[7] Bones are specified by the parent and child joints that enclose the bone and their orientation (x,y,z). For example, the Hip Left bone is enclosed by the Hip Center joint (parent) and the Hip Left joint (child) [10].

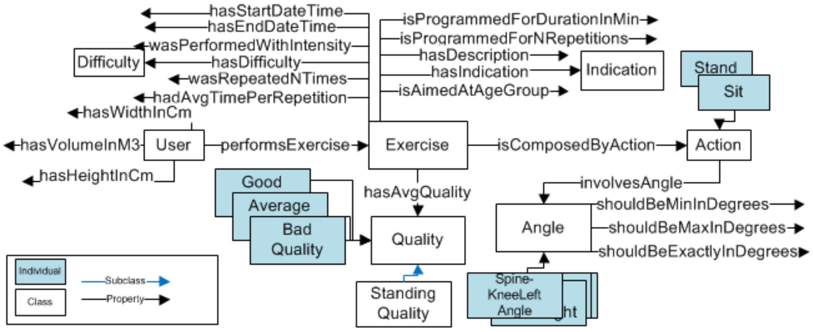

Fig. 3. Exercise & Workout Sub-Ontology

3.2 Kinect Extensions Ontology

A set of relevant classes is defined next to make sense on body, objects and actions interactions.

The class *User* identifies the person behind the Skeleton model. A user is modelled with the correspondent arms (and hands) and a set of properties that, e.g., may identify him as *PrimaryUser*[8].

Body Movement mainly represents actions executed with body limbs and articulations. Different kind of movements include to rotate, bend, extend and elevate. These can have a clockwise direction (e.g. *RotateWristClockwise*), a direction (*ElevateFootFront*), a degree or a body part to which they apply (*LeftBodyPart*).

Any physical *Object* and its properties such as dimensions, (partial) colours or number of voxels can be represented, for instance, to recognize activities such as experiments involving volume measurements. *Object actions* model interaction between objects or among users and objects thanks to Kinect Fusion API module. Examples of interactions between user and objects include to *grab, release, touch, click* etc.

The Spatial Relations Ontology [11] is reused to express physical space relations of objects as well as how they are placed or how they interact with each other, e.g. *contains, disjoint, equal and overlaps.*

Elements from NeOn Ontology engineering methodology [12] were used, e.g.: reusing ontology resources, requirements specification, development of required scenarios and dynamic ontology evolution. The main classes, data and object properties of the Kinect Ontology are presented in Table 1.

4 Ontology-based human activity reasoning

Figure 3 presents the structure of the Exercise & Workout Sub-Ontology, where the goal is to precisely model the specific movements a user performs, e.g.,

[8] Kinect Interaction layer decides which of the tracked users is primary and assigns him an ID and a primary hand, although both hands are tracked [10]

OWL Classes	**OWL Data Properties and Object Properties**
BodyMovement, BodyPart, ObjectAction, Exercise, Angle, (Image, Text, Video-)Control, Exercise(-Difficulty, Frequency, Intensity, Quality) Grammar, HandState, Indication, Location, Object, Orientation, Kinect-(Audio, Interaction, Sensor), Dictation, SpeechRecognitionEngine, TrackingMode, Bone, BoneJoint	*hasStart/EndDateTime, wasRepeatedNTimes, hadAvgTimePerRepetition, shouldBeMin/Max/ExactlyInDegrees, hasDescription, isProgrammedForNRepetitions, IsProgrammedForDurationInMin, hasCoordinateX/Y/Z, hasHeightInCm hasDifficulty, hasIndication, hasAvgQuality, performsExercise, isComposedByAction, involvesAngle, hasOrientation, hasSourceLocation, interactsWith, detectsKinectAudio, hasLoadedGrammar, hasActiveTrackingMode, detectsInteraction/Object, activatesControl, hasBoneHierarchy, isLocatedIn, hasSpatialRelation, hasInteractionMode, hasArm/Hand, hasSpeechRecognition, representsUser*

Table 1. Kinect Ontology Classes, Data and Object Properties (partial)

through the exercise duration, repetitions and quality or intensity (*Low, Medium, High*) performed.

In order to model human activities and behaviours, the state of environment variables and body postures can be abstracted so that identifying changes of interest is possible. Since existing statistical methods have demonstrated to be robust in activity monitoring [13], the Kinect ontology is intended to support these by adding context-awareness to the end-user application. For instance, long-term queries could be done, since having semantic knowledge adds the capability of integration with other sensor information, allowing for user-customization of the smart environment. Therefore, we focus on representing simple, higher level actions (lay down, washing hands, etc.) and facilitating the finding of longer term changes. Examples of the ontology in use are:

Example 1: Defining basic movement (*Stand, BendDown, TwistRight, MoveObject*, etc.) can be mapped to OWL 2, e.g., the Action *Sitting*, would be of the form:
$performsAction(Natalia, Sit) \wedge hasStartDatetime(Sit, T)$.

Example 2: When defining an activity, e.g. *Sit_StandExercise* workout, the amount of series done in a given time as well as the exercise quality can be measured. These values can be predefined according to medical parameters, e.g., the difficulty faced when sitting/standing as well as the stretching of the back when standing:
$\forall U\,, \forall Sit\_StandEx \in Sit\_Stand - Ex\,, \forall V\,: performsExercise(User, Sit\_StandEx) \wedge isComposedByAction(Sit\_StandEx, (Sit \wedge Stand) \wedge involvesAngle(Stand, LowerUpper-BackAngle) \wedge hadAngleValue(LowerUpperBackAngle, V) \wedge V < 175 \rightarrow hasAvgQuality(Sit\_StandEx, BadQuality)$.

Example 3: Historic analysis can be provided through measurements performed while doing certain activity, to monitor posture quality. E.g., having the back less straight than a year ago could be notified to make the user aware of his posture habits:

$\forall\, Stand\,, LowerUpperBackAngle1, LowerUpperBackAngle2\,, \forall\, V1, V2, D1, D2 : performsAction(Natalia, Stand) \wedge involvesAngle(Stand, LowerUpperBackAngle1) \wedge hasValue(Lower{-}UpperBackAngle1, V1) \wedge hasDateTime(LowerUpperBackAngle, D1) \wedge involvesAngle(Stand, LowerUpperBackAngle2) \wedge hasValue(LowerUpperBackAngle2, V2) \wedge hasDateTime(LowerUpperBackAngle2, D2) \wedge ((V1 - V2) > 5) \wedge T2 == (T1 + (\text{X1-XX-XX})) \wedge hasPhone(Natalias, P) \rightarrow SendSMS(P, \text{"Your back is not as extended as a year ago"})$.

Example 4: An office worker can be notified when he is not having straight back and neck:
$\forall Sit\,, \forall NeckUpperBackAngle\,, \forall V : isCurrently(Natalia, Sit) \wedge isInLocation(Natalia, NataliasOffice) \wedge involvesAngle(Sit, NeckUpperBackAngle) \wedge hadAngleValue (NeckUpperBackAngle, V) \wedge V < 175 \wedge hasPhone(Natalia, NataliasPhone) \rightarrow SendDoubleVibrationAlarm(NataliasPhone, \text{"Bad posture!"})$.

Or when he has been sitting for too long:
$\forall\, T : executesAction(Robin, Sit) \wedge hasEndDateTime(Sit, T) \wedge ((Time.Now - T) > 2h) \rightarrow sendSMS(RobinsPhone, \text{"Stand up and stretch legs!"})$.

The integration with other physiological data such as heart rate, sleep quality or stress, from sensors such as accelerometers, can be as well integrated for more complete assessments of every day functions or tasks.

5 Conclusions and Future work

We developed a OWL 2 ontology ($\mathcal{ALC}$ DL expressivity) composed of 164 classes, 53 object properties, 58 data properties and 93 individuals, based on the Kinect for Windows API. The structure of the ontology is based on Kinect Natural User Interface, Kinect Interaction, Fusion and Audio modules.

We believe that ontologies can and should play a vital part in this development to help abstracting atomic gestures for an incremental, fine, and coarse grained activity recognition. In this way, automatic reasoning for inferring of novel information is facilitated. For instance, the *Exercise* & *Workout* ontology classes allow the data integration and registration to follow the evolution of a person's performance through the quality of the workout or rehabilitation program. We exemplified the usage of the proposed ontology with different domain examples.

In future work, there is an imminent need to conduct evaluation experiments with the ontology developed, for validating its modelling accuracy, as well as its reliability. We are convinced that an appropriate combination of computer vision algorithms with semantic models of human movement and interaction can significantly improve context-awareness, recognition accuracy and activity analysis precision.

Another interesting research direction for the future is the use of fuzzy ontologies, which offer ways for dealing with vague and imprecise data. During the modelling process of the Kinect ontology, we found features susceptible to be modelled in an imprecise way. Due to the imprecision inherent to the environment, these features could benefit from being expressed through an extension of the ontology to Fuzzy OWL 2. This would ease the *looseness* of the model and facilitate user interaction, as linguistic labels can be used for natural language-based customization.

6 Acknowledgements

The work presented was funded by TUCS (Turku Centre for Computer Science).

References

1. Weiser, M.: The computer for the twenty-first century. Scientific American **165** (1991) 94–104
2. Gómez-Romero, J., Patricio, M.A., García, J., Molina, J.M.: Ontology-based context representation and reasoning for object tracking and scene interpretation in video. Expert Systems with Applications **38**(6) (June 2011) 7494–7510
3. Chen, L., Nugent, C.D.: Ontology-based activity recognition in intelligent pervasive environments. International Journal of Web Information Systems (IJWIS) **5**(4) (2009) 410–430
4. Kim, E., Helal, S., Cook, D.: Human activity recognition and pattern discovery. Pervasive Computing, IEEE **9**(1) (jan.-march 2010) 48 –53
5. Henry, P., Krainin, M., Herbst, E., Ren, X., 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) (2012) 647–663
6. Kinect@Home Project, KTH: `http://www.kinectathome.com` (2012)
7. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. UIST '11 (2011) 559–568
8. d'Aquin, M., Noy, N.F.: Where to publish and find ontologies? a survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web **11**(0) (2012) 96 – 111
9. Foukarakis, M.: Informational system for managing photos and spatial information using sensors, ontologies and semantic maps. PhD thesis, Technical University of Crete. (2009)
10. Kinect for Windows: `http://www.microsoft.com/en-us/kinectforwindows/develop/`
11. Hudelot, C., Atif, J., Bloch, I.: Fuzzy spatial relation ontology for image interpretation. Fuzzy Sets Syst. **159**(15) (August 2008) 1929–1951
12. Suárez-Figueroa, M.C.: NeOn Methodology for building ontology networks: specification, scheduling and reuse. PhD thesis, Universidad Politécnica de Madrid (2010)
13. Ismail, A.A., Florea, A.M.: Multimodal indoor tracking of a single elder in an AAL environment. In: 5th International Symposium on Ambient Intelligence. (2013) 137–145