Applying Action Attribute Class Validation to Improve Human Activity Recognition

David Tahmoush
US Army Research Laboratory
2800 Powder Mill Rd, Adelphi MD
david.tahmoush@us.army.mil

Abstract

When learning a new classifier, poor quality training data can significantly degrade performance. Applying selection conditions to the training data can prevent mislabeled, noisy, or damaged data from skewing the classifier. We extend a set of action attributes and apply training case attribute selection conditions to a challenging action recognition dataset.

Short-range 3D imagers produce three-dimensional point cloud movies which can be analyzed for structure and motion information like actions. We skeletonize the human point cloud to try to estimate the joint motion, and this produces a significant number of errors as well as damaged and misrepresented cases. By selectively pruning the training cases using the extended action attributes, we improve the classifier performance on some classes by over 5% and improve on the state-of-the-art from 85% accuracy to over 88%.

In addition, discovering attribute inconsistencies in the subject actions has provided a reason behind the consistently disappointing performance of multiple algorithms upon the same data.

1. Introduction

Action recognition has utilized low-level features over higher-level class attributes and ontologies because they traditionally have been more effective. In fact, at a recent Action Classification Challenge at the 2013 international conference on computer vision (ICCV) there were no contestants who used the compiled class attributes, and the top six contestants used only low-level attributes [1]. However, an approach with class-based attributes may improve the accuracy of methods when traditional low-level features do not perform as well as expected.

One of the cases when low-level attributes do not work well is when the data is mislabeled or corrupted in some way. The worst case is mislabeling, where the classifier is trained on the wrong data. Given enough accurate data, mislabeling of a few cases is not a severe handicap, but on smaller datasets it can make a significant difference. We isolated one action dataset that had been of particular trouble for low-level approaches, the Action Set 2 from the MSR-Action3D dataset [2] which has eight action classes with sixty instances of each class. The dataset is compiled using a time-of-flight depth sensor to capture the human actions. The 3-D depth maps are skeletonized to reduce the dimensionality of the data. This skeletonization procedure has some failures and some mistakes. How-ever, the skeletons readily provide ontologically significant action attributes, such as 'Body Part Articulation-Arm = One_Arm_Motion' which can be used to validate the skeletonization.

Instead of learning an attribute ontology from the data, we apply a knowledge-based approach and utilize a previously developed model of the attributes and skeletal relationships for particular activities. However, we needed to extend the model to characterize the particular activities in the dataset. In order to apply a previously developed human action ontology to the skeletonized actions, we utilized and extended the action attributes compiled for the Action Classification Challenge [1] to fit the different actions in the MSR-Action3D dataset [2]. An excerpt of some of the attributes for a punching action is shown in Table 1.

By applying the attribute ontology to the training set, we could prevent inaccurate, inconsistent, damaged, or mislabeled cases from damaging the classifier without reducing the richness of the features. There is the risk of that an excessively strict ontology would dramatically reduce the size of the training set and result in performance degradation, so a preliminary classification should be per-formed and the effect of applying knowledge-based ontological restrictions to the training set can be tested and the final result calculated on an evaluation dataset. A subset of the action attribute ontology is shown in Table 1. Using the knowledge-based ontology, we could identify skeletonization mistakes that made certain cases ontologically inconsistent.

The next section discusses related work on action recognition and the application of ontologies to action

recognition. Section 3 describes the dataset. Section 4 describes the ontology, while section 5 describes the classifier that is improved using a knowledge-based ontology. Section 6 tabulates the results and Section 7 contains the conclusions.

Table 1. A subset of the action attributes that are present or absent in the action 'Punch'

Action Attributes	Punch
Body Part Articulation-Arm = One_Arm_Motion	1
Body Part Articulation-Arm = Two_Arms_Motion	1
Body Part Articulation-Head = Facing_Up	0
Body Part Articulation-Head = Facing_Front	1
Body Part Articulation-Head = Facing_Sideways	1
Body Part Articulation-Head = Straight_Position	1
Body Part Articulation-Head = Tilted_Position	1
Body Part Articulation-Torso = Down_Forward_Motion	0
Body Part Articulation-Torso = Twist_Motion	1
Body Part Articulation-Torso = Bent_Position	1
Body Part Articulation-Torso = Straight_Up_Position	1
Body Part Articulation-Feet = Touching_Ground	1
Body Part Articulation-Feet = In_Air	0

2. Related Work

Human activity recognition has already been explored using images and video. Activity recognition techniques can be grouped into data-driven [3] and knowledge-driven [4] approaches. Data-driven techniques use machine learning approaches to discern an activity from the training data. Space-time methods such as space-time volumes, spatio-temporal features, and trajectories have been successful. For classification, generic approaches like support vector machines and hidden Markov models have done well.

Space-time approaches treat video as spatial (x,y) and temporal (t) axes [5-10]. An action can be described as a 3D shape in space-time and compared to labeled actions [11,12] with extensions [5-7]. For video, the spatial dimensions are x and y. By including the z value from depth images, reasonable recognition can be achieved [9]. Improvement can be achieved in recognition accuracy by including the third spatial dimension [6]. An action graph approach on a bag-of-3D-points can encode actions and produce an improvement in recognition accuracy [10]. Random occupancy patterns (ROP) from the 4D video volume were tested on MSR-Action3D dataset, and achieved 86.2% accuracy [13]. Three-dimensional joint positions were used to develop a view-invariant posture representation [14]. Combining features from RGB images, depth maps, and skeleton joints to recognize human activities has also been done [15].

HMMs are often applied for classification of activities with real-time performance. Body joints can be obtained [15] and the temporal patterns of the joint feature vectors were tested on the MSR-Action3D dataset [2], achieving 88.2% recognition accuracy [17] across all classes. Naïve-Bayes-Nearest-Neighbor was also used to achieve similar accuracy, [18] demonstrating that many classifiers would work given a thoughtful choice of features. The same work observed that the actionlet mining method was effective to handle noise and errors in skeleton joint positions. Recurrent neural networks [19], dynamic temporal warping [20], and hidden Markov models have also been used [21].

An action ontology provides a description of the activity using well-structured terminology with a number of properties that are measureable. A well-built ontology could be used, understood, and shared between humans and computers [22-24]. A human action ontology was developed [1] that was used and extended for this work.

3. Dataset

The MSR-Action3D dataset [2] is a benchmark dataset for 3D action recognition that provides sequences of depth maps and skeleton joints. It includes 20 actions performed by 10 subjects performing each action 3 times. An example is shown in Figure 1 for "high wave", where the motion of the arms, legs, head, and torso are shown with the depth dimension removed.

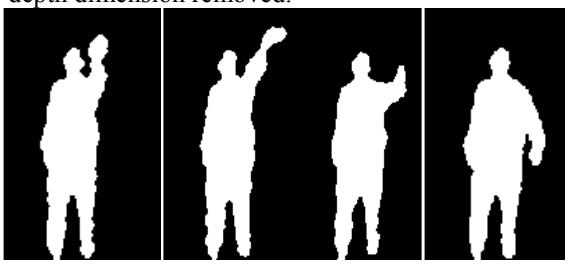

Figure 1: Example action of "high wave" from the MSR-Action3D dataset.

On datasets like MSR-Action3D, the inter-class variability is going to be somewhat small due to the similarity of the activities such as "hammer" and "forward punch" as well as the presence of composite activities such as "pickup and throw" versus "overhand throw" and versus "bend over". Since the activities to recognize are very similar to each other, the classifier will have to

handle this robustly. In many approaches to this dataset, it is the composite action of "pickup and throw" which is often misclassified because most of the motion is identical to the "bend over" action or the "throw" action. In order to perform better classifications, many authors have broken the dataset into groups as shown in Table 2. Although this reduces the number of potential confuser actions, many similar pairings are still present in the subgroups. Researchers have had particular difficulty with Action Set 2, having an average of 6-10% reduction in performance.

Table 2. Two subsets of actions used for the MSR-Action3D dataset

Action set 1 (AS1)	Action set 2 (AS2)
Horizontal wave (2)	High wave (1)
Hammer (3)	Hand catch (4)
Forward punch (5)	Draw x (7)
High throw (6)	Draw circle (9)
Hand clap (10)	Two hand wave (11)
Bend (13)	Forward kick (14)
Tennis serve (18)	Side boxing (12)
Pickup throw (20)	Draw tick (8)

The activities in this dataset are defined by pose, such as "forward punch" and "side boxing" defined to be separate classes of "punch" as well as "forward kick and "side kick" which are only defined in the reference frame of the individual's pose. Since the activities are defined in pose-relative terminology, the data to utilize should be extractable in the subject's pose reference frame.

4. Ontology and Attributes for Actions

An ontology for action recognition [1] was evaluated for the smaller set of actions Action Set 2 of the MSR-Action3D dataset [2]. Attributes that were useful in separating classes were used, like 'Body Motion = Twisting' which only is in the majority of 'Side Boxing' cases and 'Body Part Articulation-Arm = Two Arms Motion' which only is in the majority of 'Two Hand Wave' cases. We also ex-tended the ontology to include 'Body Part Articulation-Arm = One Arm Raised Head Level' and 'Body Part Articulation-Arm = One Arm Extend Side' to help evaluate the 'Side Boxing' and 'High Wave' classes.

The implementation of heuristics to test for ontological aberrations in the training data was simplified by utilizing the extracted skeletal joints. For example, testing for the percentage of time that the 'Body Part Articulation-Arm = One Arm Raised Head Level' was a simple check on whether the location of the hand was above the location of the neck. The check was whether $[(P_{13}(3) > P_3(3)\ \&\ P_{12}(3) < P_3(3)) \mid (P_{12}(3) > P_3(3)\ \&\ P_{13}(3) < P_3(3))]$ was true for an extended percentage of the time.

We found that the MSR-Action3D dataset is well-suited for developing quick checks on attributes and then applying ontological reasoning. Some of the heuristic tests that discovered aberrations from the extracted skeletons are shown in Table 3. 'Body Motion = Twisting' found that one subject would consistently twist his body to the side when performing 'One Hand Catch' (A4S2E1-3) which was inconsistent with the class attributes for the rest of the subjects and is shown in Figure 2. 'Body Motion = Twisting' also found that 2 of 10 subjects did not perform measurable twisting when performing 'Side Boxing'.

The attribute 'Body Part Articulation-Arm = One Arm Raised Over Head' found that two cases of 'Side Boxing' kept their hand above their head for an extended period of time (A12S4E1 and A12S3E3) which was inconsistent. Attribute 'Body Part Articulation-Arm = Two Arms Raised Over Head' found that one case of 'High Wave' used both hands (A1S2E3) which could create confusion with 'Two Hand Wave' as shown in Figure 3. 'Body Part Articulation-Arm = One Arm Open To The Side' found that one case of 'Side Boxing' did not actually extend the arm to the side (A12S2E1) as shown in Figure 4. We extended the attribute 'Body Part Articulation-Leg = Up Down Motion' to create "Body Part Articulation-Hand = Up Down Motion' which found that two cases (A7S2E1 and A7S6E3) of 'Draw X' and one case (A8S6E3) of 'Draw Tick' had significantly more up and down hand motion than the rest of the cases.

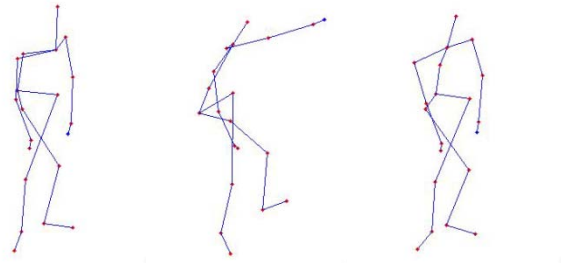

Figure 2: Incompatible motion for A4S2E1 meant to be 'Hand Catch'.

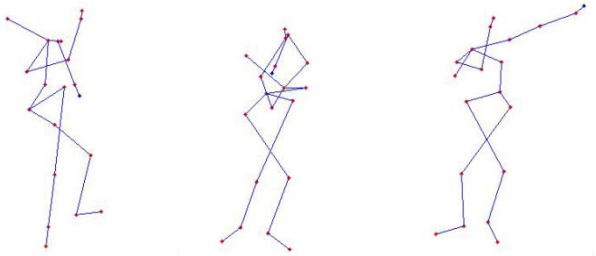

Figure 3: Damaged skeletal extractions for A1S2E3 meant to be 'High Wave' often has two hands in the air.

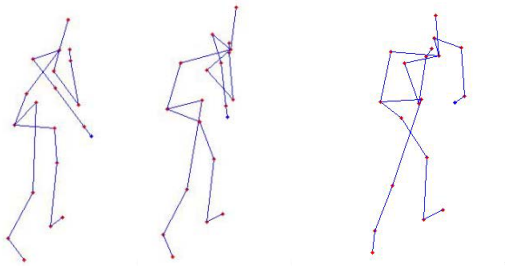

Figure 4: Damaged skeletal extractions for A12S2E1 meant to be 'Side Boxing' does not extend his arm for most of the action.

Some of the aberrant cases found were poor skeletal extractions while others were subjects interpreting the action differently or repeating the action multiple times.

5. Classification Approach

The 3D joint velocities are used to recognize the motion of the human body. Our key observation is that representing the human movement as joint velocities results in effective features that can be represented in the pose reference frame and which are invariant under time dilation. In other words, motions done more slowly or quickly can be robustly compared by time-dilating the extracted velocities using Dynamic Time Warping [25]. In some sense, the actions can be stretched in the time domain to compare to other actions.

For this work, 20 joint positions are tracked using a skeleton tracker [15] and compiled into a time series of joints i depicted as $\boldsymbol{p}_i(t) = (x_i(t); y_i(t); z_i(t))$ at a frame t. The coordinates are then normalized to reduce dependencies on height, initial body orientation and location. Examples of the extracted skeletons are shown in Figures 2, 3, and 4.

For each joint i, we extract the normalized pose-referenced velocity features by taking the difference between the position of joint $i(t)$ and that of joint $i(t-1)$ and dividing by the time step dt. The resulting velocity vector is:

$$\mathbf{v}i(t) = ((xi(t)-xi(t-1))/dt; (yi(t)-yi(t-1))/dt; (zi(t)-zi(t-1))/dt) \quad (1)$$

for each joint i in the skeletonized action. Note that the normalization has left the extracted velocities in the initial pose reference frame of the subject, which means that the extracted velocities should be able to distinguish pose-defined activities. The actions can be visualized by looking at the velocities over time. There are 20 joint velocities with three dimensions each tracked over time as shown in a Figures 5 and 6 for subjects 2 and 3, respectively.

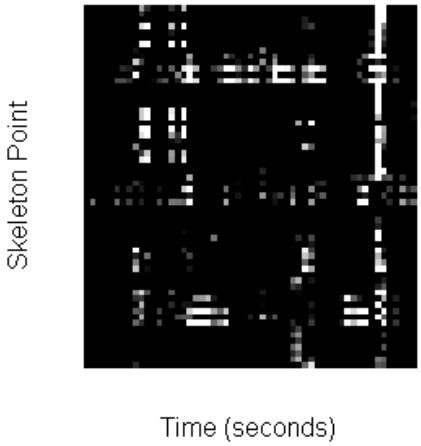

Figure 5: The joint-velocity magnitude heatmap for high wave by subject 2.

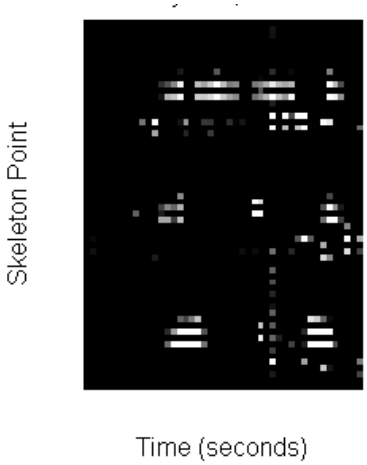

Figure 6: The joint-velocity magnitude heatmap for high wave by subject 3.

The images in Figure 5 and 6 show the log of the absolute value of the joint-velocity. The total collection time for each signature is different, with the collect on subject 2 lasting longer than the collect on subject 3.

We did not try to pick out the relevant features using PCA as had been done in other work [26] though this may be done in future work for comparison. Since the other work kept 85% of the eigenvalues, the dimensionality reduction is not great, and the eigenvectors may be less intuitive than the original skeletal joint velocities.

To classify a signature, the joint-velocity heatmap is compared to a known database of joint-velocity heatmaps. The comparison utilizes a Dynamic Time Warping [25] of the joint-velocity heatmaps as a distance function between signatures. A large measured distance between signals is a measure of the difference between signals, while a small measured distance is indicative of the similarity of two signals.

The classification of a test action with the database is performed using the smallest distance in a nearest neighbors approach. The advantage of this approach is that

the most similar action in the database will be selected as the probable for the test case. When there is high intra-class variability in a small database, finding multiple cases for k-nearest neighbors may not work well. However, a nearest neighbors classifier can be highly sensitive to mislabeled training cases as one bad training case can propagate out to many test cases. To reduce the influence of poor training data, an action ontology is applied to prune damaged or inaccurate cases from the training data.

6. Experimental Results

Several initial experiments were run on the MSR-Action3D dataset without ontological improvement to the training set. The initial results on Action Set 2 is shown in a confusion matrix in Figure 7. The results are similar to other published work, as shown in Table 3. The refinement using an action ontology has not yet been applied, but already the classification is above 87%.

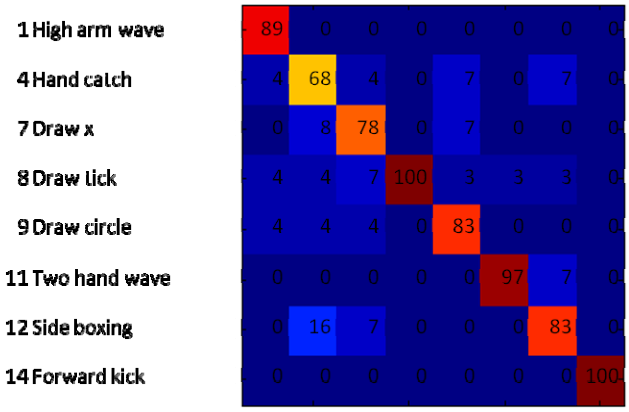

Figure 7: The initial results on action set 2. The predicted class number is in the y-axis while the actual class is on the x-axis.

Table 3. Performance versus method for Action Set 2 cross-subject MSR dataset

Method	Accuracy
Li et al [4]	71.9%
Lu et al [11]	85.5%
Yang et al [15]	84.1%
Chen et al [16]	83.3%
Initial Method	87.2%
Initial Method with Ontology	88.6%

One reason that this approach might work well is shown in Figure 8. The similarity metric between subjects performing the same action is remarkably bad for most cases, but for one case it does surprisingly well. This signature variability may help explain the challenges on this dataset.

The classification of a test action with the database is performed using the smallest distance in a nearest neighbors approach. The advantage of this approach is that the most similar action in the database will be selected as the probable classifier for the test case. When there is high intra-class variability in a small database, finding multiple cases for k-nearest neighbors may not work well, while a simple nearest neighbors will find the correct class. However, a nearest neighbors classifier can be highly sensitive to mislabeled training cases as one bad training case can propagate out to many test cases. To reduce the influence of poor training data, an action ontology is applied to prune damaged or inaccurate cases from the training data.

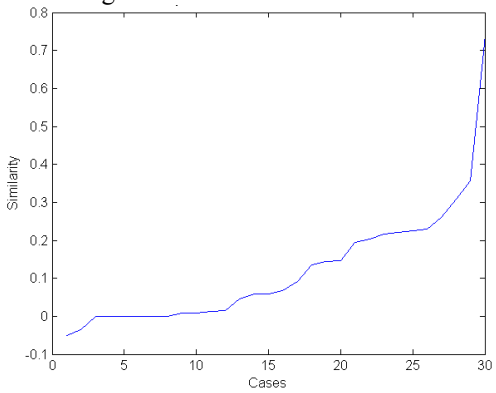

Figure 8: The similarity measure between 'High Throw' A1S1E1 and the rest of the 'High Throw' cases. This case is correctly classified.

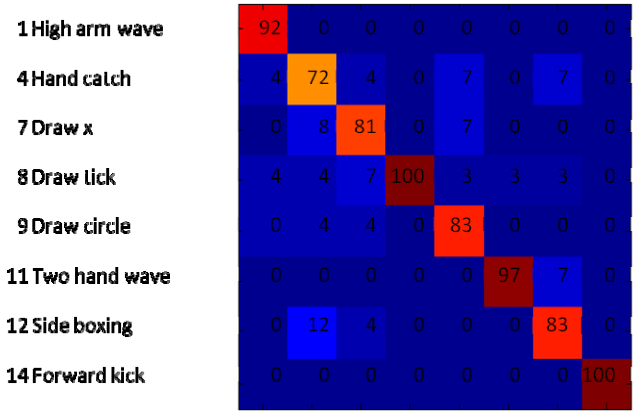

Figure 9: The results on action set with ontological training class validation. The predicted class number is in the y-axis while the actual class is on the x-axis.

The performance of the classification after the ontological pruning is done to the training data is shown in Figure 9. Slightly better performance can be achieved by utilizing the ontology to isolate and prune poor examples from the training set. However, those damaged examples

are still present in the test set since we are using leave-one-subject out cross-validation.

In addition to improving classification, a better understanding of the limitations of the dataset have been acquired. The recognition of invalid or damaged training set data gives additional information on why those cases are incorrectly classified when they are used as test cases in cross-fold validation. For example, case A4S2E1 is misclassified using this approach. This case will probably be misclassified in every approach because it has attributes that are nearly identical to the attributes of side boxing by other subjects, with both twisting of the body and an arm extending out to the side. Since many methods have been applied to this particular dataset and all have performed relatively poorly on this particular set of actions, it may be more of a dataset issue than a classifier issue.

The performance in the literature on this particular set of actions from the MSR Action dataset has been consistently lower than other action sets from the same database [2, 14, 18, 26]. The attribute inconsistencies in multiple actions of this subset found by this work may be a large part of the explanation of why presumably sound methods may be failing.

7. Conclusions

This work has shown that recognition of human actions can be improved using ontological models to remove inconsistent, mislabeled, or damaged data from the classifier training set. By incorporating human knowledge about action attributes into the ontology, cases which are inconsistent can be pruned from the training set.

However, this work also showed additional information about the dataset. By applying ontologically based heuristics to the data we are able to isolate reasons why performance is limited. Of special note was the discovery of the twisting motion that was expected in the 'Side Boxing' class but was only there 80% of the time, and it was not expected in the 'Hand Catch' class but was discovered in 10% of the data. The inconsistent attribute link between classes helps explain why there exist significant cross terms in the confusion matrix between the two classes in the work of multiple researchers.

References

[1] Jiang, Y.G., Liu, J., Zamir, A.R., Laptev, I., Piccardi, M., Shah, M., & Sukthankar, R.: THUMOS challenge: action recognition with a large number of classes. http://crcv.ucf.edu/ICCV13-Action-Workshop/ (2013).

[2] Li, W., Zhang, Z., & Liu, Z.: Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW),* 9-14 (2010, June).

[3] Ye, J., Dobson, S., & McKeever, S.: Situation identification techniques in pervasive computing: A review. *Pervasive and mobile computing*, *8*(1), 36-66 (2012).

[4] Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z.: Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(6), 790-808 (2012).

[5] Roh, M. C., Shin, H. K., & Lee, S. W.: View-independent human action recognition with volume motion template on single stereo camera. *Pattern Recognition Letters*, *31*(7), 639-647 (2010).

[6] Ni, B., Wang, G., & Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision,* 193-208 (2013).

[7] Wu, D., Zhu, F., & Shao, L.: One shot learning gesture recognition from rgbd images. In *Computer Vision and Pattern Recognition Workshops (CVPRW),* 7-12 (2012, June).

[8] Li, W., Zhang, Z., & Liu, Z.: Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW),* 9-14 (2010, June).

[9] Zhang, H., & Parker, L. E.: 4-dimensional local spatio-temporal features for human activity recognition. In *International Conference on Intelligent Robots and Systems,* 2044-2049 (2011, September).

[10] Malgireddy, M. R., Inwogu, I., & Govindaraju, V. A temporal Bayesian model for classifying, detecting and localizing activities in video sequences. In *Computer Vision and Pattern Recognition Workshops (CVPRW),* 43-48 (2012, June).

[11] Bobick, A. F., & Davis, J. W.: The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *23*(3), 257-267 (2001).

[12] Ahad, M. A. R., Tan, J. K., Kim, H., & Ishikawa, S.: Motion history image: its variants and applications. *Machine Vision and Applications*, *23*(2), 255-281 (2012).

[13] Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y.: Robust 3d action recognition with random occupancy patterns. In *European Conference on Computer Vision,* 872-885 (2012).a

[14] Xia, L., Chen, C. C., & Aggarwal, J. K.: View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW),* 20-27 (2012, June).

[15] Sung, J., Ponce, C., Selman, B., & Saxena, A.: Unstructured human activity detection from rgbd images. In *International Conference on Robotics and Automation,* 842-849 (2012, May).

[16] Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., & Blake, A.: Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(12), 2821-2840 (2013).

[17] Wang, J., Liu, Z., Wu, Y., & Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR),* 1290-1297 (2012, June).

[18] Yang, X., & Tian, Y. (2014). Effective 3D action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, *25*(1), 2-11.

[19] Martens, J., & Sutskever, I.: Learning recurrent neural networks with hessian-free optimization. In *International Conference on Machine Learning,* 1033-1040 (2011).

[20] Müller, M., & Röder, T.: Motion templates for automatic classification and retrieval of motion capture data. In *ACM SIGGRAPH/Eurographics symposium on Computer animation*, 137-146 (2006, September).

[21] Lv, F., & Nevatia, R.: Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European Conference on Computer Vision,* 359-372 (2006).

[22] Gu, T., Wang, X. H., Pung, H. K., & Zhang, D. Q.: An ontology-based context model in intelligent environments. In *Communication networks and distributed systems modeling and simulation conference*, 270-275 (2004, January).

[23] Riboni, D., & Bettini, C.: Context-aware activity recognition through a combination of ontological and statistical reasoning. In *Ubiquitous Intelligence and Computing*, 39-53 (2009).

[24] Chen, L., Nugent, C. D., & Wang, H.: A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering, 24*(6), 961-974 (2012).

[25] Berndt, D. J., & Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series. In KDD workshop 10(16), 359-370 (1994, July).

[26] Chen, C., Liu, K., & Kehtarnavaz, N.: Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*, 1-9 (2013).