Proc. of IEEE International Conference
on Intelligent Robots and Systems, San Francisco, CA, 2011.

4-Dimensional Local Spatio-Temporal Features for Human Activity Recognition

Hao Zhang and Lynne E. Parker

***Abstract*— Recognizing human activities from common color image sequences faces many challenges, such as complex backgrounds, camera motion, and illumination changes. In this paper, we propose a new 4-dimensional (4D) local spatio-temporal feature that combines both intensity and depth information. The feature detector applies separate filters along the 3D spatial dimensions and the 1D temporal dimension to detect a feature point. The feature descriptor then computes and concatenates the intensity and depth gradients within a 4D hyper cuboid, which is centered at the detected feature point, as a feature. For recognizing human activities, Latent Dirichlet Allocation with Gibbs sampling is used as the classifier. Experiments are performed on a newly created database that contains six human activities, each with 33 samples with complex variations. Experimental results demonstrate the promising performance of the proposed features for the task of human activity recognition.**

I. Introduction

Human activity recognition has played an important role in applications such as security, surveillance, smart homes and human-machine interface. Especially in robotics, the ability of a robot to understand the activity of its human peers is critical for the robot to collaborate effectively and efficiently with humans in a peer-to-peer human-robot team. However, recognizing human activities from sequences of color images is a very challenging problem due to complex backgrounds, illumination changes, camera motion, variations of human appearance and diversity of human activities.

In our work, we focus on developing 4-dimensional local spatio-temporal features, and applying these features to identify human activities from a sequence of RGB-D images, i.e., color images with depth information. Our work is motivated by the recent success of "bag of features" representation for recognizing objects and human activities [1] from a sequence of images. Based on the plausible assumption that a global human activity can be characterized by the local motions and therefore by spatio-temporal features, the "bag of features" representation models an activity as a distribution of the spatio-temporal features that are computed from the color image sequences.

A robust human activity recognition system can use not only intensity information, but also depth information. This provides a reliable way to separate humans from the environment, providing possibilities for overcoming the problems caused by complex backgrounds and camera motion. Thanks to the emergence of Kinect, an affordable color-depth camera developed by Microsoft, it is faster and easier to obtain color and depth information. In our application, a Kinect sensor is installed on top of a Pioneer 3DX robot, as shown in Figure 1. Kinect consists of an RGB camera to obtain color information and a depth camera to obtain depth information. For the depth camera, the IR emitter projects an irregular pattern of infrared dots with varying intensities, and the depth sensor reconstructs a depth image by recognizing the distortion in this pattern.

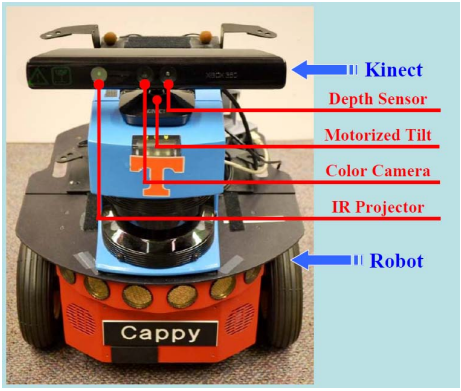


Fig. 1: Installation of Kinect sensor on a Pioneer 3DX robot.

In this paper, we propose a new 4D local spatio-temporal feature that combines both intensity and depth information, and apply this feature to identify not only repetitive but also sequential activities and activities with small motions. To our knowledge, no 4D local spatio-temporal features have been developed to address the problem of recognizing human activities using a RGB-D camera.

The rest of the paper is organized as follows. Section II gives a concise review of the existing work on feature extraction from a sequence of images. Section III describes the proposed 4D local spatio-temporal features that are extracted from intensity and depth videos. Section IV introduces the classifier used for activity recognition. Section V presents the test results on a newly created database. Finally, Section VI concludes the paper and indicates future work.

II. Related Work

A RGB-D camera or multiple cameras are often employed for 3D visual data acquisition, such as a sequence of realtime RGB-D images or recorded multi-view videos. A major step involved in an activity recognition system is the extraction of low-level features from 3D visual data, which always consist of massive amounts of raw information in the form of

This paper is based in part upon work supported by the National Science Foundation Grant No. 0812117.

H. Zhang and L. E. Parker are with the Distributed Intelligence Laboratory, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996-3450, {haozhang, leparker}@utk.edu

spatio-temporal pixel variations. But most of the information, like background clusters and colors of human clothes, is not directly relevant for identifying the activities in the visual data. Thus, feature extraction from raw 3D visual data is of great necessity and importance to get useful information. Although most previous work on feature extraction focused on using 2D videos [2], several approaches to extract features from 3D videos have been proposed in the past few years.

A simple technique is to apply a 3D centroid trajectory as features to identify human activities in 3D visual data, in which a human is represented as a point that indicates the 3D location of the human in the visual data [3]. In general, the feature of centroid trajectory is suitable for representing a human that occupies a small region in an image. Another method to extract features in 3D visual data relies on human shape information, such as a history of 3D human silhouette [4]. A third type of technique to detect features for human activity recognition is on the basis of 3D human models, such as a 3D human skeleton model [5] or a 3D articulated body-part model [6]. The robustness of the features on the basis of 3D human shape and body models relies heavily on the performance of foreground human segmentation and body part tracking, which are hard-to-solve problems due to dynamic background and occlusions.

The precursors to the 4D features proposed in this paper are the local spatio-temporal features extracted from 2D visual data [7], which have recently become a popular activity representation and have shown promising performance for the task of human activity recognition. A spatio-temporal feature represents some local texture and motion variations regardless of the global human appearance and activity. A global activity is presented as a bag of local spatio-temporal features. Dollar *et al.* [8] extracted such features using separable filters in the spatial and temporal dimensions. Laptev *et al.* [9] detected the features on the basis of a generalized Harris corner detector with a set of the spatio-temporal Gaussian derivative filters. Other spatio-temporal features are also proposed based on the extended Hessian saliency measure [10], a salient region detector [11], or global information [12]. A detailed evaluation of several spatio-temporal features in [7] indicates their similar performances for the task of human activity recognition.

III. Proposed 4D Spatio-Temporal Features

In this work, we propose the 4D local spatio-temporal feature as a representation of human activities, which combine both intensity and depth information obtained from the Kinect sensor. Our work is inspired by the local features developed by Dollar [8].

A. Preprocessing

Both the color camera and the depth camera in the Kinect sensor are first calibrated to obtain their intrinsic parameters and accurately map between depth pixels and color pixels. Then, a first order approximation is applied for converting the raw 11-bit disparity value to an 8-bit depth value to form a depth image. To reduce computational complexity, the color and depth images are resized to a resolution of 320 by 240. The color image is then converted to an intensity image, and histogram equalization is employed to reduce the influence of illumination variation. For depth images that are very noisy, erosion and dilation are used to remove noise and small structures, and then hole filling is performed.

B. Feature Detection

To extract feature points from the preprocessed sequences of intensity and depth images, a response function is computed at each pixel using both intensity and depth information within a hyper 4D cuboid that is illustrated in Figure 2. A feature point is also detected, which corresponds to a local maximum of the response function. The location of a feature is determined on both the intensity image sequence $I(\boldsymbol{x}, t)$ and the depth image sequence $D(\boldsymbol{x}, t)$. To exploit spatial correlation, spatial filters are applied on all intensity and depth images:

$$I_s(\boldsymbol{x}_o, t) \;=\; \big(I(\boldsymbol{x}, t) \circ f(\boldsymbol{x}, t|\delta)\big) * p(\boldsymbol{x}|\sigma)\big|_{\boldsymbol{x}=\boldsymbol{x}_o} \quad (1)$$

$$D_s(\boldsymbol{x}_o, t) \;=\; \big(D(\boldsymbol{x}, t) \circ f(\boldsymbol{x}, t|\delta)\big) * p(\boldsymbol{x}|\sigma)\big|_{\boldsymbol{x}=\boldsymbol{x}_o} \quad (2)$$

where '$*$' denotes convolution, '$\circ$' denotes Hadamard product (entry-wise matrix multiplication), $\boldsymbol{x} = \{x, y\}$ is a pixel, $\boldsymbol{x}_o$ is the current pixel, and $f(\boldsymbol{x}, t|\delta)$ is an indicator function parameterized by δ. The parameter δ controls the spatial scale along the depth dimension:

$$f(\boldsymbol{x}, t) = \mathbf{1}(\big|D(\boldsymbol{x}, t) - D(\boldsymbol{x}_o, t)\big| \leqslant \delta) \quad (3)$$

and $p(\boldsymbol{x}|\sigma)$ is a 2D Gaussian filter applied along the spatial dimensions x and y. The parameter σ of the Gaussian filter controls the spatial scale along x and y dimensions:

$$p(\boldsymbol{x}|\sigma) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}}. \quad (4)$$

A temporal filter is also applied along dimension t on the spatial-filtered sequences:

$$I_{st}(\boldsymbol{x}_o, t) \;=\; I_s(\boldsymbol{x}_o, t) * g(t|\tau, \omega)\big|_{t=t_o} \quad (5)$$

$$D_{st}(\boldsymbol{x}_o, t) \;=\; D_s(\boldsymbol{x}_o, t) * g(t|\tau, \omega)\big|_{t=t_o} \quad (6)$$

where $g(t|\tau, \omega)$ is a 1D complex-value Gabor filter given by:

$$g(t|\tau, \omega) = \frac{1}{\sqrt{2\pi}\tau} \cdot e^{-\frac{t^2}{2\tau^2}} \cdot e^{i(2\pi\omega t)} \quad (7)$$

where τ controls the temporal scale of the detector, and in all cases we use $\omega = 3/\tau$.

Finally, the response strength of pixel $\boldsymbol{x}_o$ at time t_o can be computed by the response function:

$$R(\boldsymbol{x}_o) = \alpha \cdot \|I_{st}(\boldsymbol{x}_o)\|^2 + (1 - \alpha) \cdot \|D_{st}(\boldsymbol{x}_o)\|^2 \quad (8)$$

where α is a mixture weight.

Any region undergoing an observable motion can induce response. Each local maximum in the response function is detected as a feature point where significant motion occurs. Instances of response images that are computed with the response function are shown in Figure 3. Notice that a feature point is not detected in a complete 4D space, because the

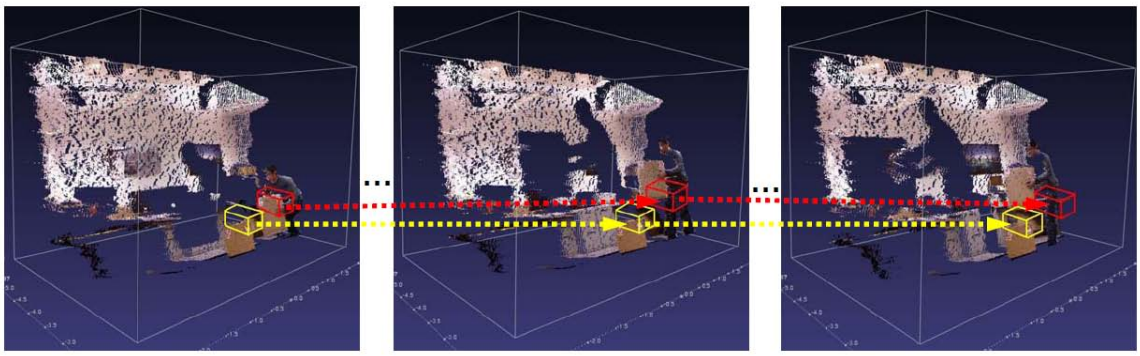


Fig. 2: An illustration of the 4D spatio-temporal hyper cuboids for feature detection and description. In the spatial dimensions, a 3D cuboid, depicted with the cube, is placed at each pixel for feature detector and at each detected feature point for feature descriptor. Then, a 1D temporal filter, depicted with a dotted arrow, is used to connect all the 3D spatial cuboids at the same pixel into a 4D hyper cuboid. This procedure makes the hyper cuboid contain both space and time information.

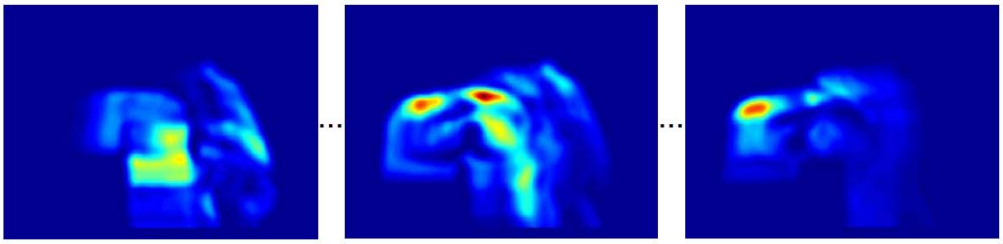


Fig. 3: Response images that are computed with the response function using the color and depth information in Figure 2.

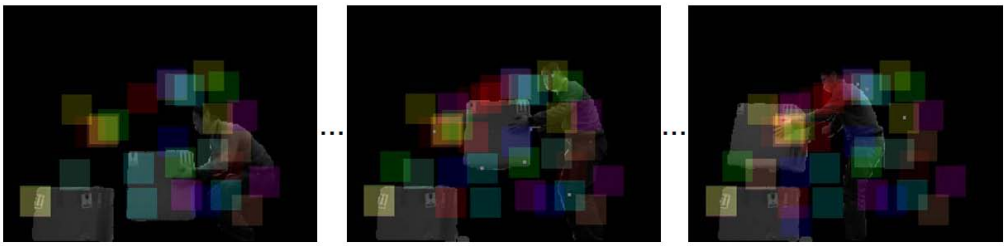


Fig. 4: Actual spatio-temporal cuboids that are extracted from the intensity and depth information in Figure 2. Each cuboid is projected to the 2D foreground intensity images for an easier display.

depth value is actually a function of x and y; not all 3D points $\{x, y, z\}$ have an intensity value. But the depth $D(\boldsymbol{x})$ still provides useful information along the z dimension.

C. Feature Description

For the feature descriptor, a hyper 4D cuboid is centered at each feature point $\{x, y, z, t\}$. The size of the hyper cuboid is $\{2s\sigma, 2s\sigma, 2s\delta, s\tau\}$, where s is the side-length ratio of the descriptor and detector cuboid. Instances of the extracted cuboids are depicted in Figure 4. To get a descriptor for each 4D hyper cuboid, the intensity and depth gradients along x, y, and t dimensions are computed. The computed gradients from both intensity and depth pixels are concatenated to form a feature vector. The size of the feature vector equals the number of pixels in the cuboid times the value of the time scalar times the number of the gradients directions times 2 (for intensity and depth values). A feature vector often contains over 10^5 double elements. Therefore, in general, the features are intractable.

To solve this problem, the principal component analysis (PCA) is applied. PCA is a dimensionality reduction method, which projects each feature vector to a lower dimensional space. To obtain a more compact representation of the feature vector, a k-means algorithm with Euclidean distance is used to cluster a large number of feature vectors computed from the training data. A spatio-temporal codeword is then defined to be the center of a cluster, and the codebook is defined to be the set that contains all the codewords. Thus, each extracted feature vector can be assigned to a codeword, and each video sequence can be represented as a bag of codewords from the codebook. It has been noted in [1] that clustering is useful to handle the feature that contains the patterns of scale change and camera motion, as long as the feature is not extremely different from the features used to form the codebook.

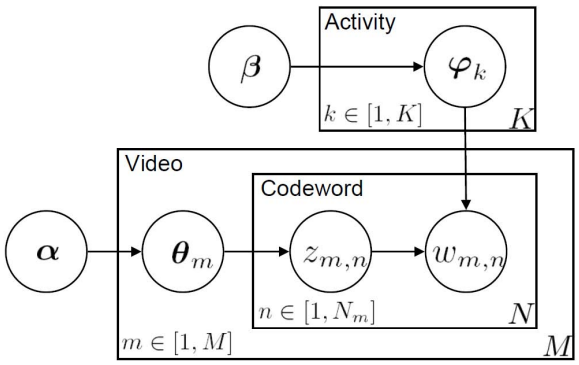


Fig. 5: Graphical representation of LDA model. The boxes are plates, representing replicates.

IV. Graphical Model for Activity Recognition

In this work, we use Latent Dirichlet Allocation (LDA) to categorize human activities, which is first introduced by Blei [13] for text classification. LDA is a generative probabilistic model, which includes priors in a Bayesian manner to combine domain knowledge and avoid overfitting, and provides an unsupervised learning framework to perform meaningful reasoning[1]. LDA models each activity as a distribution over codewords and each video as a distribution over activities, which allows a video to be explained by the latent activities. The graphical representation of LDA is shown in Figure 5. Suppose we have a set of M videos that record K activities, and the size of the codebook is V. The generative process of LDA is illustrated in Figure 6, where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are the Dirichlet parameters.

For the task of human activity recognition using LDA, the major problem is to estimate and infer the parameter $\boldsymbol{\theta}_m$, i.e., the distribution of activities for video m. However, exact parameter estimation and inference is intractable. To address this issue, several approximate methods have been proposed such as variational methods [13], Gibbs sampling [14], and expectation-propagation [15]. Gibbs sampling is a Markov chain Monte Carlo (MCMC) method, which often yields a relatively efficient algorithm for approximate estimation and inference in high-dimensional models such as LDA [16]. Therefore, we use Gibbs sampling to approximately estimate and infer the parameters of LDA in our work. One can refer to these papers for a better understanding of Gibbs sampling. Here we only show the most important formula. Let $\boldsymbol{w}$ and $\boldsymbol{z}$ be the vectors of all codewords and their activity assignments in the entire set of videos. Then, the activity assignment of a particular codeword t is sampled from the multinomial distribution using Gibbs sampling:

$$p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w}) = \frac{n^{(t)}_{k,\neg i} + \beta_t}{[\sum_{v=1}^{V} n^{(v)}_k + \beta_v] - 1} \cdot \frac{n^{(k)}_{m,\neg i} + \alpha_k}{[\sum_{j=1}^{K} n^{(j)}_m + \alpha_j] - 1} \quad (9)$$

[1]Although LDA is used as a classifier in our work, in general, the 4D local spatio-temporal feature does not rely on any specific classifier.

• ACTIVITY PLATE
for each activity $k \in [1, K]$ **do**
  Choose the per-activity codeword proportions:
    $\boldsymbol{\varphi}_k \sim \text{Dirichlet}(\boldsymbol{\beta})$
end for
• VIDEO PLATE
for each video $m \in [1, M]$ **do**
  1) Choose the number of codewords:
    $N_m \sim \text{Poisson}(\xi)$
  2) Choose the per-video activity proportions:
    $\boldsymbol{\theta}_m \sim \text{Dirichlet}(\boldsymbol{\alpha})$
  • CODEWORD PLATE
  for each codeword $n \in [1, N_m]$ in video m **do**
    1) Choose the per-word activity assignment:
      $z_{m,n} \sim \text{Multinomial}(\boldsymbol{\theta})$;
    2) Choose the spatio-temporal codeword:
      $w_{m,n} \sim \text{Multinomial}(\boldsymbol{\varphi}_{z_{m,n}})$;
  end for
end for

Fig. 6: Generative process for LDA

where $n^{(t)}_{k,\neg i}$ is the times the codeword t is assigned to activity k except the current assignment, and $n^{(k)}_{m,\neg i}$ is the number of codewords in video m that are assigned to activity k except the current assignment. After Gibbs sampling is complete, each element in the parameters $\boldsymbol{\theta}_m$ can be estimated as:

$$\theta_{m,k} = \frac{n^{(k)}_m + \alpha_k}{\sum_{j=1}^{K} n^{(j)}_m + \alpha_j}. \quad (10)$$

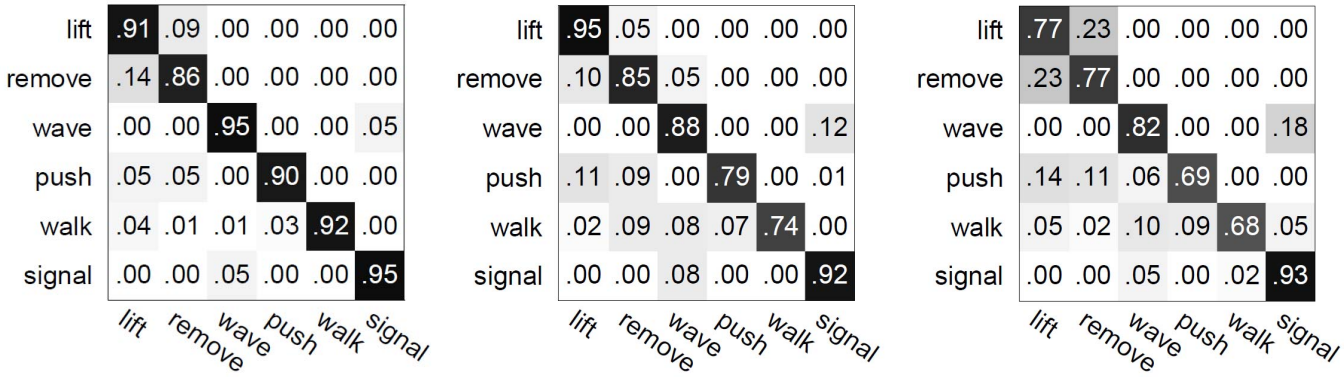
V. Experimental Results

To evaluate the performance of the proposed 4D local spatio-temporal features for the task of activity recognition, we establish a database whose data are collected with a Kinect that is installed on a Pioneer mobile robot, as shown in Figure 1. During the data collection, the tilt of the Kinect sensor is adjusted approximately 10 degrees upward to avoid recording the floor. We also considered illumination change, dynamic background and variations in human motions when recording the data to make them more challenging and more interesting. The database currently contains six types of human activities, including two sequential activities ("lifting" and "removing"), three repetitive activities ("pushing", "waving", and "walking"), and one activity with small motion ("signaling"). Each activity has 33 samples, with each sample lasting 2 to 5 seconds. Each sample consists of two parts: a color video and a calibrated depth video. The instances of the activities are illustrated in Figure 7. The color and depth frames are depicted by the first and second rows respectively, and the third row gives an intuitive 3D view of the activities. To our knowledge, no such database exists before our work.

We divide the database into three groups. Each group contains all activities with each activity having 11 samples. We detect the feature point and describe the corresponding

(a) Lifting (b) Removing (c) Waving (d) Pushing (e) Walking (f) Signaling

Fig. 7: Example image sequences from videos in our database. The database contains 6 types of human activities: lifting, removing, pushing, waving, walking and signaling. The database contains 33 samples for each activity, which are collected from the Kinect sensor on a Pioneer robot in the environments with complex variations.



(a) Intensity&Depth (Average accuracy = 91.50%) (b) Depth (Average accuracy = 85.50%) (c) Intensity (Average accuracy = 77.67%)

Fig. 8: Confusion matrices with different information using LDA model with a codebook of size 600. Rows represent actual classes, and columns represent predicted classes.

4D spatio-temporal cuboid with the procedure described in Section III. The parameters of the feature detector are set to $\sigma = 5$, $\delta = 255$, $\tau = 3$, and $\alpha = 0.5$. Each 4D spatio-temporal cuboid is then described with a feature descriptor of its intensity and depth gradients. The feature descriptor is then projected to a lower-dimensional feature vector with 60 elements using PCA. In order to build the codebook that is used for LDA, all feature vectors from the training data are clustered into 600 clusters using k-means. Each cluster is then indexed by a codeword.

We compare our 4D local spatio-temporal features extracted using both intensity and depth information to the features using either intensity or depth information that is introduced by Dollar [8]. Because of the limited amount of data, we use the leave-one-out testing paradigm to obtain a performance estimation of our methods for the task of activity recognition; i.e., for each run, one group is selected as the training set, and the remaining groups are used as the testing sets. Because Gibbs sampling for approximate parameter estimation and inference of LDA has some random components, the experiment results are reported as the average over 20 runs.

Under these settings, we learn and recognize human activities using the LDA model with the Gibbs sampling technique. The confusion matrix using the proposed 4D local spatio-temporal features is given in Figure 8a. Each column of the confusion matrix corresponds to the predicted category, and each row corresponds to the ground truth class. Using the 4D spatio-temporal features, an average accuracy of 91.50% is achieved. The confusion matrix shows that the largest confusion lies between "lifting" and "removing" within the category of sequential activities. This is consistent with the "bag-of-codewords" representation that assumes each codeword is independent of others. Thereby, this rep-

resentation loses the information of the relative positions of the codewords in a frame. On the other hand, it should be noted that the proposed feature has captured some time information, which enables the classifier that has no ability to model time series, such as the LDA model, to recognize different sequential activities correctly in most cases.

The confusion matrices of activity recognition using the LDA model with features that are extracted from either depth information or intensity information are illustrated in Figure 8b and Figure 8c, respectively. The LDA model gets an accuracy of 85.50% with local features extracted from depth video and an accuracy of 77.67% with features from intensity video, which indicates that the depth information is more important than the intensity information for our database. A possible explanation is that the color videos recorded in the home environment suffer significant illumination variations that are caused by weather changes, as illustrated in Figure 7d, 7e and 7f. In the office environment, the computer monitors lead to a dynamic background, which also distract the feature detector from detecting useful human motions. But the depth sensor is not sensitive to either the illumination variations or the dynamic background, as long as the target human is far enough from the background.

In general, for the task of human activity recognition using the LDA model with the Gibbs sampling technique, the proposed 4D local spatio-temporal features outperforms the features using only intensity or depth information. On the other hand, these features also exhibit some similar patterns, which can be observed from their confusion matrices. First, all three features can model time series to some extent. But the locality of these features causes them to lose the position information between features, leading to a moderate accuracy of identifying sequential activities. Furthermore, the activities such as "pushing" and "walking", in which the human crosses the entire horizontal field of view of the sensor, are often confused by several other activities. For instance, the activity "pushing" is confused by "lifting" and "removing", and the activity "walking" is often confused by all the other activities. This phenomenon can be explained partially that "pushing" and "walking" contain some basic motions in the task of box-pushing, such as holding a box and moving the body, which will lead to some similar features exhibited by other activities. Finally, "waving" and "signaling" are in general only confused by each other due to their similarity that only human arms move as a human performs these activities.

VI. Conclusion

In this paper, a new 4D local spatio-temporal feature is proposed using both intensity and depth information, and the relative feature detector and descriptor are analyzed. Then, the features are used for the task of activity recognition with the LDA model as a classifier. To estimate the performance of the proposed features, a new database is created for the task of human activity recognition. Each sample in the database consists of a color video and a calibrated depth video, which are collected from Kinect installed on a Pioneer robot.

The experimental results on the database show that the proposed 4D local spatio-temporal features extracted from the intensity and depth videos outperform the local features extracted using only intensity or depth video. The experimental results also indicate that in the case of significant illumination variations and dynamic background, the depth information is more important than the intensity information for detecting more relevant features.

Future work includes developing more sophisticated descriptors with the ability to adjust the size of the 4D hyper cuboid adaptively to deal with the scale variations. We also plan to combine our feature detector with some human detecting techniques to extract more relevant features and ignore the features from the background environment. Other aspects of our future work include improving the classifiers based on the bag-of-features representation and enlarging the database with more activities and more samples. In the long term, we plan to let a moving robot gain the ability to recognize human activities.

References

[1] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, pp. 299–318, Sept. 2008.

[2] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] O. Brdiczka, M. Langet, J. Maisonnasse, and J. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Trans. Autom. Sci. Eng.*, vol. 6, pp. 588–597, Oct. 2009.

[4] P. Yan, S. Khan, and M. Shah, "Learning 4D action feature models for arbitrary view action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–7, Jun. 2008.

[5] J. Y. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," *AAAI Wksp. on Pattern, Activity and Intent Recognition*, to be published.

[6] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3D human body tracking with an articulated 3D body model," in *IEEE Int'l. Conf. on Robotics and Automation*, pp. 1686–1691, May 2006.

[7] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, Sept. 2009.

[8] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE Int'l Wksp. on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Oct. 2005.

[9] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, pp. 107–123, Sept. 2005.

[10] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conf. on Computer Vision*, pp. 650–663, 2008.

[11] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Trans. Syst. Man Cybern.*, vol. 36, pp. 710–719, Jun. 2005.

[12] S. F. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," *IEEE Int'l Conf. on Computer Vision*, pp. 1–8, 2007.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[15] T. Minka and J. Lafferty, "Expectation-Propagation for the generative aspect model," in *18th Conf. on Uncertainty in Artificial Intelligence*, pp. 352–359, 2002.

[16] G. Heinrich, "Parameter estimation for text analysis," *Technical Report, University of Leipzig*, 2005.