A Novel Multimodal Data Analytic Scheme for Human Activity Recognition

Girija Chetty[1] and Mohammad Yamin[2]

[1] University of Canberra, Australia
[2] Department of MIS, King Abdulaziz University, Saudi Arabia
Girija.Chetty@canberra.edu.au, myamin@kau.edu.sa

Abstract. In this article, we propose a novel multimodal data analytics scheme for human activity recognition. Traditional data analysis schemes for activity recognition using heterogeneous sensor network setups for eHealth application scenarios are usually a heuristic process, involving underlying domain knowledge. Relying on such explicit knowledge is problematic when aiming to create automatic, unsupervised or semi-supervised monitoring and tracking of different activities, and detection of abnormal events. Experiments on a publicly available OPPORTUNITY activity recognition database from UCI machine learning repository demonstrates the potential of our approach to address next generation unsupervised automatic classification and detection approaches for remote activity recognition for novel, eHealth application scenarios, such as monitoring and tracking of elderly, disabled and those with special needs.

Keywords: Multimodal, PCA, LDA, RBM, Activity recognition, Feature learning.

1 Introduction

Automatic human activity recognition for complex eHealth application scenarios requiring unsupervised monitoring and tracking of elderly, disabled and those with special needs, is a very challenging problem, especially when data is captured remotely using heterogeneous sensor networks, with sensors capturing the data related to activities being performed by humans and objects in the environment. We investigate the potential of multimodal machine learning and data mining methods for discovering learning features for human activity recognition using heterogeneous sensor networks with humans and object in the environment.

Over the last few years, recognizing activity from motion sensors and accelerometer sensor data patterns has become a popular area of research in ubiquitous computing and computer vision area, and one of the most successful applications of image analysis and understanding. There is an urgent need for development of automatic activity recognition systems from such heterogeneous sensor data, for visualising the goal of a next-generation automatic surveillance technology for health care of elderly and disabled, with applicability to development

K. Liu et al. (Eds.): ICISO 2014, IFIP AICT 426, pp. 449–458, 2014.

of remotely instrumented home care environments. Several physiological and bio-mechanical studies have shown that most of the human activity in performing day-to-day activities is inherently multimodal, and is based on kinematic interaction between several motion articulators, involving lower and upper body parts and other bio-mechanics of joints. It is person specific based on body weight, height, joint mobility in the lower and upper body, and type of activity being performed and the objects in the environment. For an automatic recognition of an activity the human is performing, there is a need to take into consideration multimodal cues available from the human body parts, from the surrounding environment and from other objects present in the environment.

If automatic activity recognition systems can be built based on this concept, it will be a great contribution to eHealth area, particularly for remote activity monitoring and recognition using heterogeneous sensor networks in aged care and disability care sector. However, each of these cues or traits captured from heterogeneous wireless sensors on their own are not powerful enough for ascertaining activity: a combination or fusion of each of them, along with an automatic processing technique can result in robust activity recognition. In this article, we propose usage of a publicly available activity recognition dataset, and use of novel multimodal techniques based on semi-supervised machine learning for automatic activity recognition. It is to be noted, that since user cooperation is not mandatory upon data collection, this novel strategy is suitable for monitoring the elderly and disabled for remote home care monitoring scenarios.

In this article, we propose the use of a principled approach involving feature extraction techniques based on automatic semi-supervised discovery, such as principle component analysis (PCA) and linear discriminant analysis (LDA), and novel deep learning approach. Further, we propose that the score level fusion of these features can enhance the performance of activity recognition scheme as compared to single mode image features. Fusing features captured from heterogeneous sensors from the sensor network at the score level is more effective than fusion at feature level, as the incompatible, asynchronous sensors in the sensor network can be combined using different fusion rules in a synergistic manner[2]. The experimental evaluation of the proposed approach with a publicly available activity recognition database [1] shows a significant improvement in recognition performance as compared to other methods proposed in the literature. Rest of the article is organised as follows. Next Section describes the background and motivation for proposed work, followed by the proposed multimodal activity recognition scheme in Section 3. The details of the experiments performed are described in Section 4, and conclusions and plans for further work are described in Section 5.

2 Background

Activity recognition is an essential requirement for automatic monitoring of elderly disabled, and those with special needs, for next generation automated home care environments. In general, sensors, which are either worn on the body and/or

embedded into objects and the environment, are utilized to capture aspects of movement or a human's behavior. Ideally, by applying data analysis, image and signal processing and pattern classification techniques, this sensor data can be automatically analyzed yielding a real-time classification of the activities that users (patients, humans who are aged or those who have special needs) are engaged in. Activity recognition can be considered a classical (multi-variate) time series or sequence analysis problem, for which the task is to detect and classify those contiguous portions of sensor data streams that cover activities of interest for the target application. The predominant approach to activity recognition is based on a sliding window procedure, where a fixed length analysis window is shifted along the signal sequence for frame extraction. Subsequent frames overlap to some degree in this sliding window approach, but are usually processed separately. Preprocessing then transforms raw signal data into feature vectors, which are subjected to statistical classifiers that eventually provide activity hypotheses. As for any pattern recognition task, the keys to successful activity recognition are: (i) choice of appropriate features to be extracted from raw sensor data; and (ii) the design of suitable learning classifiers. The machine learning and data mining literature describes a wide variety of supervised machine learning approaches involving the stages of feature extraction, feature selection and learning classifiers. By contrast, comparatively little systematic research has addressed the problem of feature design, with almost all previous work using heuristically selected general measures. These features are either calculated in the time domain, calculated on symbolic representations of the sensor data, or spectra based. The lack of systematic research on appropriate features for automatic unsupervised classification or even semi-supervised classification is one of the major shortcomings of current activity recognition systems. For example, it is questionable whether the next generation of eHealth applications for remote home-care scenarios for activity monitoring of elderly and disabled, for behavioral analysis, fall or injury detection, or monitoring of vital health parameters can be realized based on the use of such heuristically selected features alone, requiring constant human/expert intervention. Such problems require intelligent unsupervised or at least semi-supervised quantitative analysis of the underlying sensor data captured, which are beyond the capabilities of current procedures.

However, recent developments in the data mining and machine learning field have the potential to overcome this shortcoming by automatically discovering novel feature representations for such activity recognition from heterogeneous sensor networks. In this article, we present a novel approach to feature extraction and investigate the suitability of feature learning for activity recognition tasks. We utilize a learning framework, which automatically discovers suitable feature representations that do not rely on application-specific feature design and engineering by human experts. We use semi-supervised feature learning techniques, namely well-known principal component analysis and linear discriminant analysis, and recently proposed deep learning technique, and show how the automatic discovery of features outperform traditional statistical and supervised learning features for an activity recognition application. Such a novel feature extraction procedure has important implications for the development of future eHealth applications such as remote monitoring of

home-care environments for elderly, disabled and those with special needs, since no manual optimization is required. The deep learning approach allows for in-depth analysis of the underlying multimodal data from different sensors, as the new representation based on semi-supervised machine learning implicitly highlights the most informative portions of the analyzed data [2, 3,4]. This is likely to be important for new classes of activity analysis such as new or anomalous activity recognition where there is no previous information available in the databases, which is normally the case with unstructured home care health environment.

Each of the sensor node, whether it is for tracking the person data or that of the other objects in the environment of the person, can contribute significantly to detecting the higher level activity being performed. While the sensor data captured from human body can be termed as primary sensor data, the sensor data captured from surrounding objects in the environment can be termed as a soft sensor data or secondary data. Soft or secondary data often captures the high level information of the environment where the human performs the activity, and though acts as weak information for recognising the human activity, does help in enhancing the robustness of activity recognition if multiple heterogeneous secondary sensor data from the environment is used in appropriate combination. [5, 6]. In other words, if we combine complementary information from another source, this multimodal combination is expected to be powerful for activity recognition. Further, use of an appropriate automatic processing scheme for processing this multimodal sensor network, can enhance the performance and robustness of the system. For example, researchers in [7, 8] have found that multi-modal scheme involving simple PCA features on combined heterogeneous sensor input results in significant improvement over single mode sensor data. In addition, other recent attempts to improve the recognition accuracy include multiple heterogeneous set of sensors has been reported in [9], [10]. The fusion of complementary sensor node information from disparate sources for activity recognition, however, did not attract much attention from the research community. This could be due to difficulty in acquiring the data, and processing and making sense out of them.

3 Dataset for Multimodal Activity Recognition

For experimental evaluation of our proposed multimodal activity recognition scheme, we used publicly available UCI OPPORTUNITY Activity Recognition Dataset [1]. The OPPORTUNITY Dataset for Human Activity Recognition from Wearable, Object, and Ambient Sensors is a dataset devised to benchmark human activity recognition algorithms. A subset of this dataset comprises the readings of motion sensors recorded while users executed typical daily activities:

- Body-worn sensors: 7 inertial measurement units, 12 3D acceleration sensors, 4 3D localization information.
- Object sensors: 12 objects with 3D acceleration and 2D rate of turn
- Ambient sensors: 13 switches and 8 3D acceleration sensors

- Recordings: 4 users, 6 runs per users. Of these, 5 are Activity of Daily Living runs characterized by a natural execution of daily activities. The 6th run is a "drill" run, where users execute a scripted sequence of activities.
- Annotations/classes: the activities of the user in the scenario are annotated on different levels: "modes of locomotion" classes; low-level actions relating 13 actions to 23 objects; 17 mid-level gesture classes; and 5 high-level activity classes.

The activity recognition environment and scenario has been designed to generate many activity primitives, yet in a realistic manner. Subjects operated in a room simulating a studio flat with a deckchair, a kitchen, doors giving access to the outside, a coffee machine, a table and a chair. Each subject was recorded in 6 different runs. Five of them, termed as activity of daily living (ADL), followed a given scenario. The remaining one, a drill run, was designed to generate a large number of activity instances. The ADL run consists of temporally unfolding situations:

- Start: lying on the deckchair, get up
- Groom: move in the room, check that all the objects are in the right places in the drawers and on shelves
- Relax: go outside and have a walk around the building
- Prepare coffee: prepare a coffee with milk and sugar using the coffee machine
- Drink coffee: take coffee sips, move around in the environment
- Prepare sandwich: include bread, cheese and salami, using the bread cutter and various knifes and plates
- Eat sandwich
- Cleanup: put objects used to original place or dish washer, cleanup the table
- Break: lie on the deckchair

The drill run consists of 20 repetitions of the following sequence of activities:

- Open then close the fridge.
- Open then close the dishwasher
- Open then close 3 drawers (at different heights)
- Open then close door 1
- Open then close door 2
- Toggle the lights on then off
- Clean the table
- Drink while standing
- Drink while seated

The annotations are done on five ‘tracks’. One track contains modes of locomotion (e.g. sitting, standing, walking). Two other tracks indicate the actions of the left and right hand (e.g. reach, grasp, release), and to which object they apply (e.g. milk, switch, door). The fourth track indicates the high level activities (e.g. prepare sandwich). As can be seen, this dataset does provide an opportunity to benchmark many automatic activity recognition algorithms, consisting of classification, (semi-) supervised machine learning, automatic segmentation, unsupervised structure discovery, data imputation, multi-modal sensor fusion, sensor network research

transfer learning, multitask learning, sensor selection, feature extraction and classifier calibration and adaptation. Our experiments involved the subset of data acquired from this large database consisting of sensor data recorded by the accelerometer attached to the right arm of the subject. We considered 10 low level activities of interest plus an unknown activity category. The acceleration data were sampled with 64Hz yielding approximately 4,200 frames. Fig. 1 shows the screen shot for the dataset we used in our experiments.

Fig. 1. Sample data from OPPORTUNITY activity recognition dataset [1]

4 Features for Multimodal Activity Recognition

To analyze the performance of different low level features and their fusion for proposed multimodal activity recognition, we performed experiments that compared the capabilities of different combinations of features extracted from sensor data streams. Further, we examined score level fusion, which means there is no requirement of having identical dimensionalities of features for objective comparisons. This stands in contrast to feature level fusion where small differences in the dimensionality of the underlying data and loss of synchronism in fusion can lead to catastrophic fusion, and can have a significant impact on the estimation procedure and hence on the capabilities of the models.

To extract the low level features from raw sensor data streams for activity recognition, we used a set of statistical measures to represent frames of contiguous multidimensional sensor data. Given the 192-dimensional analysis, frames (64 × 3) provided by our sliding window procedure, we first calculated pitch and roll values. Subsequently, for each source channel (i.e. x, y, z, pitch, and roll) we then calculated

mean, standard deviation, energy, and entropy. Together with three correlation coefficients (estimated for all combinations of the x, y, z axes) this yielded a 23-D representation of the raw signal data covered by an analysis frame.

4.1 PCA-LDA Features

Principle component analysis is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. On the other hand, the LDA attempts to model the difference between the classes of data [9,10]. PCA does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made. LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis. And in our experiment, LDA shows more promising results than PCA does. We performed experiments utilizing PCA and LDA based features where the projection sub-space is spanned by those eigenvectors that correspond to the c = 18, 23, 30, and 39 largest eigenvectors. These selections of c are justified by significant drops in the eigenvalue spectrum of the data and correspond to the selected target dimensionalities of the other approaches investigated. No significant changes in classification accuracy were observed for the four choices of c, hence we present the results for c = 30.

4.2 Deep Learning Features

Auto encoder networks have proved to be a powerful tool for the generic semi-supervised or unsupervised discovery of features [11, 12]. These aim to learn a lower-dimensional representation of input data, which produces a minimal error when used for reconstructing the original data. As an alternative to PCA or LDA based feature extraction for continuous sensor streams we employed deep learning methods for auto encoder based feature learning on sequential data. The desired representation is discovered by means of a feed-forward neural network that consists of one input layer, one output layer and an odd number of hidden layers. Every layer is fully connected to the adjacent layers and a non-linear activation function is used. The objective function during training is the reconstruction of the input data at the output layer. The auto encoder transmits a description of the input-data across each layer of the network. Since the innermost layer of the network has a lower dimensionality, the transmission of a description through this bottleneck can only be achieved as result of a meaningful encoding of the input.

This non-linear low-dimensional encoding is hence an automatically learned feature representation in an semi-supervised manner. For robust model training, we learn the layers of the auto encoder network greedily in a bottom-up procedure, by treating each pair of subsequent layers in the encoder as a Restricted Boltzmann

Machine (RBM). An RBM is a fully connected, bipartite, two-layer graphical model, which is able to generatively model data. It trains a set of stochastic binary hidden units which effectively act as low-level feature detectors. One RBM is trained for each pair of subsequent layers by treating the activation probabilities of the feature detectors of one RBM as input-data for the next. Once the stack of RBMs is trained, the generative model is unrolled to obtain the final fully initialized auto encoder network for feature learning. Different methods exist to model real-valued input units in RBMs. We employ Gaussian visible units for the first level RBM that activate binary, stochastic feature detectors (Gaussian-binary). The subsequent layers can then rely on the common binary-binary RBM. The final layer is a binary linear RBM, which effectively performs a linear projection.

During training the sensor data is processed batch-wise, where each batch ideally comprises samples from all classes in the training-set. Note that the availability of the class information is not mandatory, since we expect an unsupervised learning. RBMs can also be trained in a completely unsupervised manner. However, balancing the batches with respect to the distribution of the classes, (sort of semi-supervised training), improves the model quality since it removes the potential for artificial biases.

Auto encoder networks contain a number of free parameters, including the network topology, i.e., the number of internal layers and its dimensionalities. The optimized network layout consists of a 4-layer model with 1024 units in each hidden layer and 30 units in the top one (192-1024-1024-30). In all experiments, the first layer was trained for 100 epochs while the subsequent layers were trained for 50 epochs. To reduce biasing due to class imbalance, each batch was split equally among all classes, holding 10 samples for each.

5 Experimental Results

To evaluate the performance of the proposed multimodal scheme for activity recognition, we conducted a number of experiments to examine the performance of different features and their multimodal fusion. Sensor data was analysed by means of a sliding window procedure, extracting frames of n = 64 contiguous samples, which overlap by p = 50 percent. Feature extraction was then performed on a frame-by-frame basis. The focus of our experimental evaluation was on the capabilities of the proposed feature representations. Accordingly, we did not focus on classifier optimisation but on the features themselves. So, we selected a standard, instance-based Nearest Neighbour (NN) classifier, and applied it “as is” to all tasks. Given ground truth annotations we report the classification accuracy as percentages of correct predictions provided by the NN classifier. The experiments were performed as N =7-fold cross validations. Folds were created by randomly choosing samples from the original dataset thereby respecting fold-wise balanced distributions of all classes (i.e. activities to be recognized).

The experiments involved examining the classification accuracy for different single mode features and multimodal features (score level fusion of features)

proposed. As can be seen in Figure 2, it was possible to achieve classification accuracy between 65% to 75% for different feature and their combinations. The classification accuracy obtained was 65.2% for PCA features, 67.6% for LDA features, 70.5% for RBM features, 72.4% for score level fusion of PCA and RBM features, and 74.7% for score level fusion of LDA and RBM features. We used a weighted fusion method, where the weight for each feature is assigned based on the classification score achieved in single mode classification. This strategy allows us to achieve an adaptive fusion that can be automated in future without manual intervention.

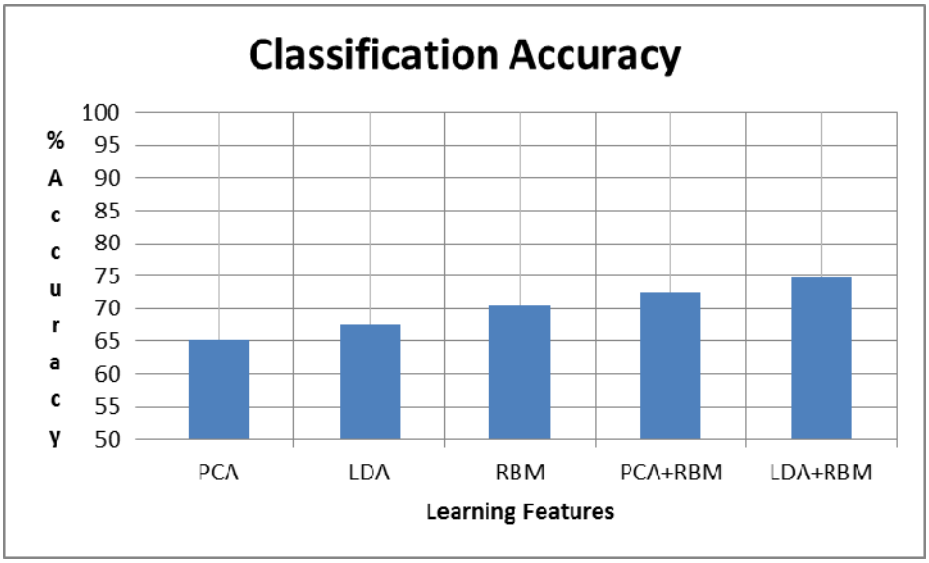

Fig. 2. Classification Accuracy for Different Learning Features

6 Conclusions and Further Plan

In this article, we proposed a novel activity recognition scheme based on multimodal fusion of semi-supervised low level features obtained from raw accelerometer sensor data. We investigated the role of simple semi-supervised subspace features which can result in development of better automatic activity recognition systems for eHealth application scenarios for monitoring the activities of elderly, disabled or those with special needs. We also examined the benefits achieved with multimodal fusion of efficient subspace features in enhancing the classification accuracy. Experimental evaluation of the proposed multimodal activity recognition scheme for a data subset from a publicly available OPPORTUNITY activity recognition UCI dataset [1], showed a significant improvement in recognition accuracy for score level fusion of features as compared to single mode features. Further research will involve investigating novel unsupervised learning approaches and combining the data from several other sensors available for activity recognition in this dataset.

References

1. Sagha, H., Digumarti, S.T., José del, R.M., Chavarriaga, R., Calatroni, A., Roggen, D., Tröster, G.: Benchmarking classification techniques using the Opportunity human activity dataset. In: IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, USA, October 9-12 (2011)
2. Huang, L.: Person Recognition By Feature Fusion. Dept. of Engineering Technology Metropolitan State College of Denver, IEEE, Denver, USA (2011)
3. Jain, A.K.: Next Generation Biometrics, Department of Computer Science & Engineering. Michigan State University, Department of Brain & Cognitive Engineering, Korea University (2009)
4. Yampolskiy, R.V., Govindaraja, V.: Taxonomy of Behavioral Biometrics. Behavioral Biometrics for Human Identification, 1–43 (2010)
5. Meraoumia, A., Chitroub, S., Bouridane, A.: Fusion of Finger-Knuckle-Print and Palmprint for an Efficient Multi-biometric System of Person Recognition. In: IEEE Communications Society subject matter experts for publication in the IEEE ICC (2011)
6. Ross, A., Jain, A.K.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
7. Chang, K., et al.: Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics. IEEE Trans. PAMI 25, 1160–1165 (2003)
8. Kittler, J., et al.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 226–239 (1998)
9. Hossain, E., Chetty, G.: Multimodal Identity Verification Based on Learning Face and Gait Cues. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part III. LNCS, vol. 7064, pp. 1–8. Springer, Heidelberg (2011)
10. Multilayer Perceptron Neural Networks, The Multilayer Perceptron Neural Network Model, `http://www.dtreg.com`
11. Hinton, G.E.: To recognize shapes, first learn to generate images. Progress in Brain Research 165, 535–547 (2007)
12. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation 18(7), 1527–1554 (2006)