DEEP FUSION OF HETEROGENEOUS SENSOR DATA

Zuozhu Liu[⋆] *Wenyu Zhang*[†] *Tony Q.S. Quek*[⋆] *Shaowei Lin*[⋆]

[⋆] Singapore University of Technology and Design
[†] Cornell University

ABSTRACT

Heterogeneous sensor data fusion is a challenging field that has gathered significant interest in recent years. In this paper, we propose a neural network-based multimodal data fusion framework named deep multimodal encoder (DME). Through our new objective function, both the intra- and inter-modal correlations of multimodal sensor data can be better exploited for recovering the missing values, and the shared representation learned can be used directly for prediction tasks. In experiments with real-world sensor data, DME shows remarkable ability for missing data imputation and new modality prediction. Compared with traditional algorithms such as kNN and Sparse-PCA, DME is more expressive, robust, and scalable to large datasets.

Index Terms— Multimodal data fusion, heterogeneous sensor data, missing data imputation, deep learning

1. INTRODUCTION

Multimodal data fusion refers to the statistical and machine-learning problem of combining data from different kinds of sensors to enable better inference, prediction and decision making [1–4]. Nowadays, wireless sensor networks are widely deployed around many domains and the sensor data collected can be used for many tasks. For instance, smart cities can make use of a variety of signals from sensors, cameras and even social media to monitor the health of the urban infrastructures and allocate resources more efficiently. Sensor data is also helpful in many other scenarios such as monitoring environmental changes, detecting infrastructural faults and improving physiological well-being [5–7].

Although sensor data is easily available, it is often incomplete due to low battery, transmission loss or faulty sensors. This incompleteness makes sensor data fusion a difficult task, degrading the accuracy of decision making and prediction tasks [8]. Traditional algorithms such as K-nearest neighbors (kNN) [9] and sophisticated dimensionality reduction techniques such as sparse-PCA [10] are widely used in many applications for missing data imputation and classification. kNN adapts local linear computations and predicts the missing data with the mean value, while sparse-PCA discovers a more reasonable representation by extracting the sparse structure of data. Both of them employ linear representations of the original data. Recently, the sparse auto-encoder was used to learn more expressive non-linear representations [11]which resulted in better performance. However, all of these methods only consider intra-modal correlations and not the inter-modal dependencies, which should be exploited to enhance fusion performance of multimodal datasets [12, 13].

In this paper, we propose the deep multimodal encoder (DME) framework based on neural network and deep learning techniques for missing data imputation and decision-making in multimodal large-scale sensor networks. Compared with kNN and Sparse-PCA, the DME is more expressive and can learn richer features because of the nonlinear activation functions. To deal with incomplete training data, we incorporate a novel learning objective function that embraces missing values and enables DME to capture latent features in the readings. To take advantage of multimodal data, the DME performs a two-stage training procedure to learn a shared representation that captures both the intra- and inter-modal correlations. Because the compressed shared representation learned by the neural network comes from a continuous transformation that preserves salient statistical properties of the raw data, new modalities may be predicted directly from the code without decompressing it.

We conduct experiments with real-world agricultural sensor data; namely, we have a large number of humidity (%), illuminance (lux) and temperature (°C) readings from multiple sensors on a high-tech farm. The results show that DME outperforms many traditional methods such as kNN, sparse-PCA and single-layer auto-encoders in missing data imputation and new modality prediction. Its ability to impute missing data degrades only slightly even when half the readings are dropped. The DME also reconstructs temperature from humidity and illuminance with an RMSE of 7°C, directly from a highly compressed (2.1%) shared representation that was learned from incomplete (80% missing) data.

The rest of the paper is structured as follows. We formulate the problem in Section 2. In Section 3, the DME framework is defined mathematically. The algorithms are evaluated in Section 4, and we conclude in Section 5.

2. PROBLEM FORMULATION

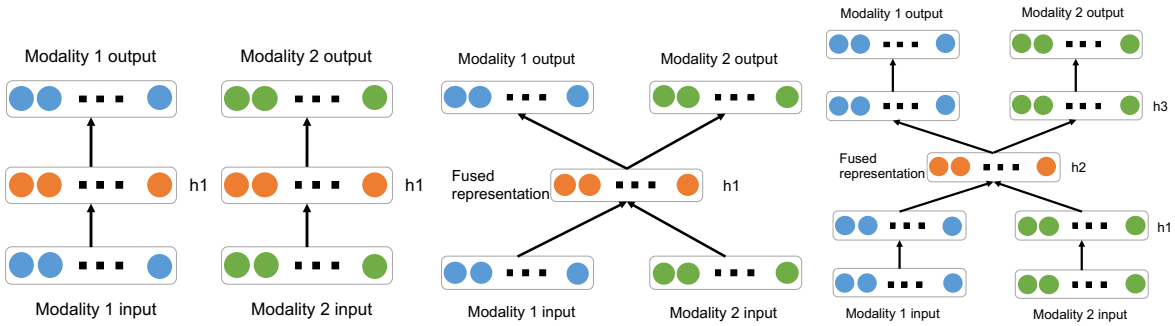
In this section, we provide the general idea behind missing data imputation in sensor datasets. For notational convenience, we use bold uppercase, bold lowercase and regular letter for matrices, vectors and scalars, respectively. $\boldsymbol{A}^\top$ and $\boldsymbol{a}^\top$ are matrix transpose and vector transpose, $\boldsymbol{A}\cdot\boldsymbol{B}$ denotes the element-wise product, $\boldsymbol{AB}$ denotes the matrix product, and $\otimes$ denotes the Kronecker product.

Suppose we have k data modality. For modality x, $\boldsymbol{X}_c = [\boldsymbol{x}^{(1)}\,\boldsymbol{x}^{(2)}\dots\boldsymbol{x}^{(N)}]^\top \in \mathbb{R}^{N\times T}$ [14] denotes the ground truth environmental matrix, where each row $\boldsymbol{x}^{(i)} \in \mathbb{R}^T$ consists of T sensor readings coming from distinct sensors or time slots. The N samples $\boldsymbol{x}^{(1)}\,\boldsymbol{x}^{(2)}\dots\boldsymbol{x}^{(N)}$ are then assumed to be independently and identically distributed. The actual incomplete environmental matrix is $\boldsymbol{X} = \boldsymbol{X}_c \cdot \boldsymbol{S}^x$, where $\boldsymbol{S}^x$ is the indicator matrix and each entry $s^x_{n,t}$ is defined as

$$s^x_{n,t} = \begin{cases} 1, \text{if } x_{n,t} \text{ is observed} \\ 0, \text{if } x_{n,t} \text{ is missing} \end{cases}$$

Suppose we fill the missing values in $\boldsymbol{X}$ to get $\hat{\boldsymbol{X}}$. Our final goal is to minimize the deviation between $\hat{\boldsymbol{X}}$ and $\boldsymbol{X}_c$:

$$\min\|(\hat{\boldsymbol{X}} - \boldsymbol{X}_c)\cdot(1-\boldsymbol{S}^x)\|^2 \tag{1}$$



(a) Uni-modal auto-encoder (b) Concatenated auto-encoder (c) Deep Multimodal Encoder

Fig. 1: Different neural network models for missing data imputation

However, $\boldsymbol{X}_c$ is not available. Hence, we change the objective function to:

$$\min \|(\hat{\boldsymbol{X}} - \boldsymbol{X}) \cdot \boldsymbol{S}^x\|^2 \tag{2}$$

where we only include the observed values [8] . The assumption here is that as long as we can reconstruct the observed values accurately, we can capture the statistical features of the sensor data, making values filled in $\hat{\boldsymbol{X}}$ reasonable approximations of the ground truth. Here we only consider one modality as an example. The multimodal case will be studied more carefully in Section 3.

In the experiments, we use the root-mean-square error (RMSE) as the metric for performance. The RMSE error e is defined as

$$e_x := \sqrt{\frac{\|(\hat{\boldsymbol{X}} - \boldsymbol{X}_c) \cdot (1 - \boldsymbol{S}^x)\|^2}{\|1 - \boldsymbol{S}^x\|}} \tag{3}$$

3. METHODOLOGY

In this section we will describe our DME model in detail. DME aims to capture both the intra- and inter-modal correlations for missing data imputation. It is a specific deep learning architecture of auto-encoders along with a novel objective function.

3.1. DME Framework

The DME framework uses auto-encoder for greedy layer-wise training as illustrated in Fig.1.(c) [15, 16]. An auto-encoder is a 3-layer neural network, consisting of T input visible units, H hidden units, and T output units, for learning representations from data. It takes N training samples, $\boldsymbol{X} = [\boldsymbol{x}^{(1)}\boldsymbol{x}^{(2)}\dots\boldsymbol{x}^{(N)}]^\top$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^T$, as input and learns parameters $\{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}\} \in \{\mathbb{R}^{T\times H}, \mathbb{R}^{H\times T}, \mathbb{R}^T, \mathbb{R}^H\}$ such that $\boldsymbol{h}^{(i)} = \boldsymbol{a}^{(2,i)} = f_1(\boldsymbol{x}^{(i)}\boldsymbol{W}^{(1)} + \boldsymbol{b}^{(1)})$ and $\hat{\boldsymbol{x}}^{(i)} = \boldsymbol{a}^{(3,i)} = f_2(\boldsymbol{a}^{(2,i)}\boldsymbol{W}^{(2)} + \boldsymbol{b}^{(2)}) \approx \boldsymbol{x}^{(i)}$ for non-linear activations f_1, f_2.

Suppose we have two modalities with incomplete environmental matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The key idea is to learn the intra-modal correlations of each modality individually in the first hidden layer before extracting the inter-modal correlations in the second hidden layer. Two traditional auto-encoder architectures, Uni-modal auto-encoder (UAE) and Concatenate auto-encoder (CAE), are showed in Fig.1.(a) and Fig.1.(b), respectively. UAE considers different modalities separately, learning only the intra-modal correlations. CAE incorporates inter-modal correlations, however, it simultaneously learns the intra- and inter-modal correlations in one layer, which prevents either of them from being learned accurately because of their vastly different statistical properties.

3.2. Missing Data Imputation

In this subsection, we describe how DME adapts to the inherent incompleteness of wireless sensor datasets. Through a new loss function in both the intra- and inter-modal learning, the missing values can be filled with the values predicted by feed-forwarding the DME.

3.2.1. Intra-modal Learning

Given two incomplete input datasets $\boldsymbol{X} \in \mathbb{R}^{N\times T_x}$ and $\boldsymbol{Y} \in \mathbb{R}^{N\times T_y}$, where N is the number of samples and T_x, T_y are respective sample dimensions, the proposed reconstruction loss of input $\boldsymbol{X}$ is

$$\tilde{J}(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{2}\| \frac{1}{\mathbb{1}^N \otimes \boldsymbol{\theta}^x} \cdot (\hat{\boldsymbol{X}} - \boldsymbol{X}) \cdot \boldsymbol{S}^x\|^2 \tag{4}$$

where $\mathbb{1}^N$ is an N dimension column vector $[1, 1, ...1]^\top$ and $\hat{\boldsymbol{X}}$ is the output of the auto-encoder. Instead of simply normalizing the loss function by the scalar $\frac{1}{N}$ in traditional auto-encoder [15], we devise a new normalizing vector $\boldsymbol{\theta}^x \in \mathbb{R}^{T_x}$ which is defined as

$$\theta_t^x = \sum_{n=1}^{N} s_{n,t}^x \ , \text{ for } t \in 1, 2, ...T_x \tag{5}$$

We can regard each entry θ_t^x as the number of observed samples in modality x of certain dimension t. The intuition is that dimensions with different number of missing entries should be weighted differently in the objective function.

To learn a better representation, we also add a novel sparsity constraint. Suppose the activation in hidden layer with H_x units is $\boldsymbol{a}^x = f(\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b})$, we denotes the mean activation as $\hat{\boldsymbol{\rho}}^x \in \mathbb{R}^{T_x \times H_x}$, which is mathematically defined as

$$\hat{\boldsymbol{\rho}}^x = [(\boldsymbol{a}^x)^\top \boldsymbol{S}^x \cdot \frac{1}{\mathbb{1}^{H_x} \otimes \boldsymbol{\theta}^x}]^\top \tag{6}$$

and the new sparsity penalty term is redefined as

$$\frac{\beta^x}{T_x}\|KL(\hat{\boldsymbol{\rho}}^x \| \rho^x)\|_1 = \frac{\beta^x}{T_x} \sum_{h=1,t=1}^{H_x, T_x} KL(\hat{\rho}_{h,t}^x \| \rho^x) \tag{7}$$

Table 1: RMSE Humidity & Temperature

	Humidity								Temperature							
Miss rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
KNN	8.31	14.56	20.98	27.90	35.41	43.56	52.36	61.8	2.45	4.66	7.12	9.81	12.71	15.83	19.13	22.70
S-PCA	15.45	17.88	19.09	21.79	26.62	33.66	42.97	54.40	6.09	7.28	7.88	8.67	10.08	12.49	15.90	19.96
UAE	5.09	5.98	6.79	7.73	8.13	9.00	10.07	11.06	2.03	2.32	2.48	2.71	2.95	3.10	3.27	3.69
CAE	**4.64**	5.98	6.40	7.36	8.05	8.87	9.75	11.07	**1.73**	2.07	2.32	2.52	2.78	**2.93**	3.28	3.81
DME	**4.64**	**5.79**	**6.37**	**7.15**	**7.85**	**8.62**	**9.50**	**10.69**	**1.73**	**2.04**	**2.29**	**2.45**	**2.68**	**2.93**	**3.14**	**3.65**

Table 2: RMSE Humidity & Illuminance

	Humidity								Illuminance							
Miss rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
KNN	8.31	14.56	20.98	27.90	35.41	43.56	52.36	61.85	**290.63**	**515.26**	762.50	1043.95	1315.64	1604.82	1887.74	2169.86
S-PCA	15.45	17.88	19.09	21.79	26.62	33.66	42.97	54.40	389.38	614.82	870.16	1081.57	1312.71	1540.43	1797.90	2069.87
UAE	5.29	5.98	6.78	7.62	8.01	8.96	10.04	11.13	683.01	715.33	753.04	793.97	833.18	890.74	959.53	**1022.86**
CAE	5.24	5.88	6.49	**7.11**	7.88	8.96	9.53	10.81	624.06	679.22	727.41	768.66	**806.65**	891.63	972.08	1095.98
DME	**4.83**	**5.73**	**6.16**	7.15	**7.45**	**8.30**	**9.31**	**10.61**	615.17	651.43	**696.10**	**739.20**	811.31	**863.50**	**955.17**	1027.21

where β^x, ρ^x is the sparse weight and predefined sparsity for the auto-encoder of modality $\boldsymbol{X}$, respectively, and $\|\cdot\|_1$ is the ℓ-1 norm. The idea here is similar to the reconstruction loss, i.e., we only enable the entries which are not missing in original input to affect sparsity in the hidden units.

Together with the unchanged decay weight regularization term, the new objective function for modality x is given by

$$\min L^x(\boldsymbol{X}) = \|\frac{1}{\mathbb{1}^N \otimes \boldsymbol{\theta}^x} \cdot (\hat{\boldsymbol{X}} - \boldsymbol{X}) \cdot \boldsymbol{S}^x\|^2 + \lambda^x \|\boldsymbol{W}^x\|^2 + \frac{\beta^x}{T_x} \sum_{h=1,t=1}^{H_x,T_x} KL(\hat{\rho}^x_{h,t} \| \rho^x) \tag{8}$$

where $\boldsymbol{W}^x$ is the weight in this intra-modal auto-encoder for modality x. In the multimodal case, we learn multiple auto-encoders for each modality individually. For instance, the final objective function for a two-modality case is defined as

$$\min L^{xy}(\boldsymbol{X}, \boldsymbol{Y}) = L^x(\boldsymbol{X}) + L^y(\boldsymbol{Y}) \tag{9}$$

where $L^y(\boldsymbol{Y})$ for modality y is defined with the same rules as $L^x(\boldsymbol{X})$. The hyper parameters, i.e., $\rho^x, \rho^y, \beta^x, \beta^y, \lambda^x, \lambda^y$, can vary among different modalities for better performance and can be found by grid search.

3.2.2. Inter-modal Learning

One main advantage of the DME is the ability to learn inter-modal correlations among different modalities. After we trained auto-encoders for each modality, we can extract the hidden-layer activations $\boldsymbol{h}^x_1, \boldsymbol{h}^y_1$ and conduct another layer-wise learning procedure to capture the inter-modal correlations. Basically, we concatenate the activations as $\boldsymbol{h} = [\boldsymbol{h}^x_1, \boldsymbol{h}^y_1]$, and train a second auto-encoder with $\boldsymbol{h}$ as the input. The objective function $L^h(\boldsymbol{h})$ we want to minimize is

$$L^h(\boldsymbol{h}) = \frac{1}{2N}\|\hat{\boldsymbol{h}} - \boldsymbol{h}\|^2 + \lambda^{xy}\|\boldsymbol{W}^{xy}\|^2 + \beta^{xy} \sum_{j=1}^{H^{xy}} KL(\rho^{xy} \| \hat{\rho}_j) \tag{10}$$

where $\hat{\boldsymbol{h}}$ is the reconstruction of $\boldsymbol{h}$, $\boldsymbol{W}^{xy}$ is weight matrix, $\hat{\rho}_j$ is the mean activation of the jth neuron, H^{xy} is the number of hidden units, and $\lambda^{xy}, \beta^{xy}, \rho^{xy}$ are the respective hyper-parameters.

In this step, by combining two modalities, we are able to mine the correlations between these modalities. This inter-modal learning procedure benefits from the intra-model correlations which were infered previously, allowing higher-order structures in the data to be captured more accurately.

4. EXPERIMENT RESULTS

4.1. Dataset

The dataset used for the experiments consists of three modalities: temperature (in degrees Celsius), humidity (relative humidity in %) and illuminance (light integral lux), collected through an agriculture sensor network of 40 sensors deployed in different locations over 4 months. After preprocessing, we get a total of 3306 samples for each modality, each consisting of 144 readings with timestamps ranging from 00:00 to 23:50 at 10-minute intervals and no data is missing. 306 samples are randomly selected for validation, 600 for testing and the remaining 2400 for training. The basic statistics of the training sets are show in Table 3.

Table 3: Dataset Statistics

	Temp.	Hum.	Illum.
Min	21.16	9.58	0
Max	60.95	100.00	98295.30
Lower Quartile	25.90	72.63	0
Median	27.60	84.42	29.68
Upper Quartile	31.28	90.87	2411.29
Standard Deviation	5.03	16.16	6635.97

4.2. Experimental Setup

We evaluate our model using the above dataset for two analytical tasks: 1) missing data imputation, 2) new modality prediction. Two different missing data imputation experiments are conducted: one for humidity and temperature (HT), and another one for humidity and illuminance (HI). The missing rate varies from 10% to 80%. To get datasets with missing values, the indicator matrix $\boldsymbol{S}$ is randomly generated with respect to the missing rate. For example, under 10% missing rate, each entry in $\boldsymbol{S}$ will be set to 0 with probability 0.1.

More formally, we are exploring the Element Random Loss pattern [8].

The parameter k in kNN is chosen as $k \approx \lceil \sqrt{n} \rceil = 49$ where n is the number of training samples. For sparse-PCA, we extract 20 sparse atoms and set the sparsity controlling parameter as 0.2. The experiments of three neural network models are conducted with Theano and the hidden layers have 6 hidden units. Mini-batch stochastic gradient descent (MSGD) is employed in optimization.

4.3. Missing Data Imputation Evaluation

Five algorithms are compared for missing data imputation, namely UAE in Fig.1.(b), CAE in Fig.1.(b), kNN [9], sparse-PCA(S-PCA) [10] and DME. The RMSE for different algorithms are shown in Table.1 and Table.2.

Humidity-Temperature: Firstly, we can observe that DME outperforms all other models . In fact, the RMSE for kNN and sparse-PCA increases dramatically as the missing rate increases. The humidity RMSE of kNN is 2 to 7 times larger than DME, and the corresponding error of sparse-PCA is also 3 to 6 times larger than DME. For temperature, DME outperforms kNN and sparse-PCA by an order of magnitude when the missing rate is greater than 50%.

We introduce another metric:

$$\text{Relative RMSE} = \text{RMSE}_{\text{Algorithm}} - \text{RMSE}_{\text{DME}} \qquad (11)$$

for further comparison. A larger relative RMSE indicates that DME is performing better. As we can see, the relative RMSE to kNN and Sparse-PCA increase significantly with the missing rate, demonstrating that DME is more robust than the linear methods, especially for high missing rates. The main reason is that kNN and sparse-PCA cannot capture the underlying data distribution accurately when lots of data is missing. However, DME employs nonlinear transformations, producing models which are much more expressive. This observation is also supported by the good performance of UAE and CAE.

Among the three neural network based models, the best performer turns out to be DME. The performance gain can be credited to the higher-order inter-modal features captured by DME. Learning such features seems to be challenging for the shallower UAE and CAE models. Another interesting observation is that when the missing rate becomes larger, the improvement also increases. The result again demonstrates that DME is more robust than the other models at learning underlying structures in the data, even when many of the values are missing.

Humidity-Illuminance: In these experiments, DME mostly outperforms other models in imputing humidity; in the few cases where it is not the state-of-the-art, the performance is only slightly worse. For illuminance which has a long-tailed distribution, DME is worse than kNN and S-PCA when the missing rate is 10% or 20% because the high variability in long-tailed distributions is not easily captured by neural networks with few hidden units. Meanwhile, when the missing rate is low, kNN is able to estimate the missing values just by querying the many neighboring sensors which do have readings. When the missing rate becomes larger, the performance of kNN degrades due to the lack of neighbors, while DME is better able to capture the diminishing information using its 6 fusion units.

4.4. New Modality Prediction

In this section, we use the fused representation of humidity and illuminance to predict temperature. We train another auto-encoder for temperature with 6 hidden units, then an over-complete feed-forward network with 12 hidden units is employed to learn a map from the extracted shared representations to the hidden activations for temperature. These three neural networks are stacked together, as shown in Fig.2. The RMSE between the predicted and the original temperature data is shown in Fig.3.

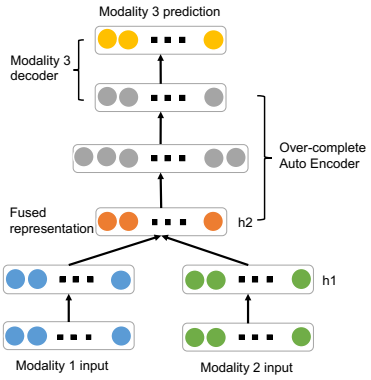


Fig. 2: DME for New Modality Prediction

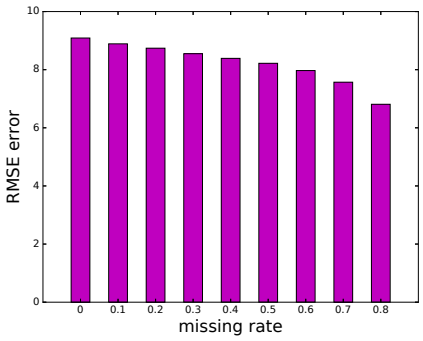


Fig. 3: New Modality Prediction RMSE

Despite missing values in the input data, our framework is still able to predict the new modality quite well, with an RMSE of less than 10°C. One counter-intuitive observation is that as the missing rate increases, the prediction RMSE decreases. Several reasons account for this phenomenon. Firstly, as showed in [17], missing values can act as a regularizer to help achieve some improvements. Secondly, we did not conduct any finetuning or hyperparameter grid search over the entire tri-modal stack to optimize the performance of the new modality prediction. These additional optimization steps could reverse the decreasing trend in the RMSE.

5. CONCLUSION

In this paper, we proposed the DME framework to overcome the challenges of missing data imputation and data fusion in wireless sensor networks. DME is able to capture both the intra- and inter-modal correlations, demonstrating outstanding performance with a real-world sensor dataset. Future work can be done on more complex dataset with different loss patterns. Together with the multi-layer hierarchical DME framework, we hope to design efficient distributed computing algorithms for sensor network that reduce power consumption, bandwidth usage and storage requirements.

6. REFERENCES

[1] David L Hall and James Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[2] Bahador Khaleghi and *et al.*, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[3] Nicolle M Correa and *et al.*, "Canonical correlation analysis for data fusion and group inferences," *Signal Processing Magazine, IEEE*, vol. 27, no. 4, pp. 39–50, 2010.

[4] Ertugrul Necdet Ciftcioglu, Aylin Yener, and Michael J Neely, "Maximizing quality of information from multiple sensor devices: The exploration vs exploitation tradeoff," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 5, pp. 883–894, 2013.

[5] Ian F Akyildiz and *et al.*, "Wireless sensor networks: a survey," *Computer networks*, vol. 38, no. 4, pp. 393–422, 2002.

[6] Mohammad *et al.* Abu Alsheikh, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 4, pp. 1996–2018, 2014.

[7] Francisco Javier Ordóñez and Daniel Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 115, Jan 2016.

[8] Linghe Kong and *et al.*, "Data loss and reconstruction in sensor networks," in *INFOCOM, 2013 Proc. IEEE*, 2013, pp. 1654–1662.

[9] Thomas M Cover and Peter E Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[10] Rodolphe Jenatton, Guillaume Obozinski, and Francis R Bach, "Structured sparse principal component analysis.," in *AISTATS*, 2010, pp. 366–373.

[11] Liang Ze Wong and *et al.*, "Imputing missing values in sensor networks using sparse data representations," in *Proc. 17th ACM Inter. Conf. on Modeling, analysis and simulation of wireless and mobile systems*, 2014, pp. 227–230.

[12] J. Ngiam and *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. on Machine Learning*, 2011, pp. 689–696.

[13] Nitish Srivastava and Ruslan R Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[14] Linghe Kong, Dawei Jiang, and Min-You Wu, "Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction," in *Distributed Computing Systems (ICDCS), 2010 IEEE 30th Inter. Conf. on*. IEEE, 2010, pp. 179–188.

[15] Andrew Ng, "http://ufldl.stanford.edu/wiki/index.php/ufldl_tutorial, ufldl tutorial," .

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] Nitish Srivastava and *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.