Modality Selection for Classification on Time-series Data

Murchana Baruah and Bonny Banerjee
Institute for Intelligent Systems, and Department of Electrical and Computer Engineering
The University of Memphis, Memphis, TN 38152, USA
{mbaruah,bbnerjee}@memphis.edu

ABSTRACT

In this paper, we investigate the optimal modality selection problem for time-series data in the context of late fusion. Multimodal emotion or action recognition is used as a testbed. Widely-used features and classifier are used for each modality drawn from five benchmark datasets. We experimented with four widely-used late fusion methods. From the classification accuracies obtained for all possible combinations of modalities in each dataset, we observe that the accuracy does not always improve with increase in number of modalities. We further show that expected information gain increases monotonically with classification accuracy in an useful interval and hence, can be used for selecting a subset of modalities for late fusion to achieve a high classification accuracy.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Emotion, action, recognition, multimodal, fusion, information gain

ACM Reference Format:
Murchana Baruah and Bonny Banerjee. 2020. Modality Selection for Classification on Time-series Data. In *MileTS '20: 6th KDD Workshop on Mining and Learning from Time Series, August 24th, 2020, San Diego, Claifornia, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Real-world time-series data is often multimodal. Learning from multiple modalities facilitates learning a richer representation which helps to make more accurate inference [7, 33]. Each modality is expected to provide unique information, otherwise it would be redundant. A challenge associated with multimodal data classification is to select a subset of modalities to maximize accuracy. This optimal modality selection problem is investigated in areas such as multimedia [2, 20, 45] and wireless sensor networks [18, 27, 36]. This problem has been explored for early fusion (see [1] for review) but rarely for decision or late fusion.

In this paper, we investigate the optimal modality selection problem for time-series data in the context of late fusion. We experimentally evaluate the effect of combining different modalities on emotion recognition accuracy using two benchmark datasets (DEAP [25], HCI Tagging [41]) containing physiological signals, and on action classification accuracy using three benchmark datasets (PAMAP2 [40], UTD MHAD [9], Berkeley MHAD [35]) containing inertial, motion capture and depth data. Several methods have been reported for action or emotion recognition using these benchmark datasets (see for example [6, 10, 25, 35, 43, 44]). However, they did not report on how to select a subset of modalities, and their experiments are limited to a few datasets. Modality selection criteria, such as independence of classifiers [26] and classifier correlation [15], have been used in late fusion models, but were not evaluated on a number of benchmark datasets.

We compare emotion or action recognition accuracy from four late fusion methods and all possible subsets of modalities from each of the five benchmark datasets. We selected candidates for the different components of a late fusion model from the literature based on their wide usage: (1) features extracted from the signal in each modality, (2) classifier, (3) late fusion methods, and (4) modality selection criterion. These candidates allow us to experiment with multiple late fusion models and draw general conclusions.

The contributions of this paper are as follows:

(1) Our experimental results reveal that, contrary to expectation, emotion or action classification accuracy does not always increase with increase in number of modalities for different late fusion methods. More data might confuse a model as data is not always informative.

(2) We empirically show that information gain increases monotonically with classification accuracy in the accuracy interval $[a, 1]\times 100\%$ where $a \in [0, 0.5]$ for two or more classes.

(3) Our experimental results show that information gain is an useful metric for selecting the optimal subset of modalities for late fusion. For the five benchmark datasets and both action and emotion recognition, subsets selected using information gain yield results comparable to the highest classification accuracy.

2 MODELS AND METHODS

Our recognition model consists of four functions: feature extraction, classification, modality selection, and late fusion (see Fig. 1). We use widely-used feature extraction, classification and fusion methods, and modality selection metrics.

Let $\{M_1, \ldots, M_n\}$ be n modalities (or signals), and x_i be the feature vector and λ_i be the classifier for modality M_i $(i = 1, \ldots, n)$. Let there be m classes, $\{\omega_1, \ldots, \omega_m\}$, such that $P(\omega_k|x_i)$ represent the posterior probability for class ω_k from classifier λ_i. [1]

MileTS '20, August 24th, 2020, San Diego, California, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

[1] Typically late fusion models, including the one in this paper, ignore temporal dependencies across modalities as they fuse modalities only after the classifiers have made their decisions. As a result, late fusion models are to some extent indifferent to the heterogeneity of and synchronicity between the different modalities. They allow modality-specific parameters (e.g., sampling rate, window length, feature vector dimension) and representation (e.g., space-time vs. frequency-time) to be chosen for each

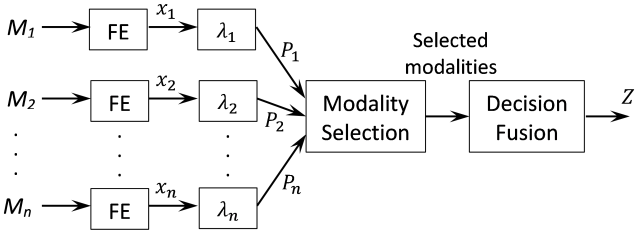

Figure 1: Block diagram for multimodal time-series classification using late fusion. 'FE' refers to feature extraction.

Definition 2.1. Correlation degree. The correlation-based classifier selection criteria for n classifiers is [15, 34]:

$$\rho_n = \frac{nN^f}{N - N^f - N^r + nN^f} \tag{1}$$

where ρ_n is the correlation degree, N^f is the number of samples misclassified by all classifiers, N^r is the number of samples classified correctly by all the classifiers, and N is the total number of samples.

Definition 2.2. Information gain. If T' denotes the predicted class labels (might be from a set of modalities after fusion or the individual modalities before fusion) and T the true class labels, the information gain of T' relative to T can be defined as [29]:

$$G(T, T') \equiv H(T) - \sum_{v \in Values(T')} \frac{|T_v|}{|T|} H(T_v) \tag{2}$$

where $Values(T')$ is the set of all possible values for T' and T_v is the subset of T for which T' has value v and $H(T)$ is the entropy of T and $\sum_{v \in Values(T')} \frac{|T_v|}{|T|} H(T_v) = H(T|T')$.

Theorem 2.1. *Expected information gain is a convex function of classification accuracy. Given the number of classes c, there exists a unique positive real number a such that the minimum of expected information gain occurs at classification accuracy a. As* $c \to \infty$, $a \to 0$.

For the trivial case $c = 1$, $a = 1$. For the nontrivial case $c = 2$, $a = 0.5$ (computed from Eq. 2). Given a particular c, for each classification accuracy $\{0, 0.1, 0.2, ..., 1\}$, we randomly generated at least 10^6 confusion matrices and computed the corresponding mean (or expected) information gain. The expected information gain as a function of classification accuracy is shown in Fig. 2 for $c = 2, 3, 7, 25$. In each case, the expected information gain is a convex function of classification accuracy. As c increases, the unique location of the minimum of this function decreases. Since classification accuracy cannot be negative, the location is lower bounded by zero.

modality independently. One way to exploit temporal dependencies across modalities in late fusion models is by learning a mapping between the representations of each pair of modalities, either directly (e.g., [30]) or via intermediate joint representations (e.g., [17]), such that the signal in each modality can be generated from the signal in another. These features can then be used for classification in each modality. Relevant topics include challenges of learning features from time series [21], generative models for learning features from time series [3, 12–14, 31, 32], generative models for learning joint representations from multimodal time series [4, 5], and opportunistic sensor selection [22, 23]. Exploiting temporal dependencies across modalities is beyond the scope of this paper.

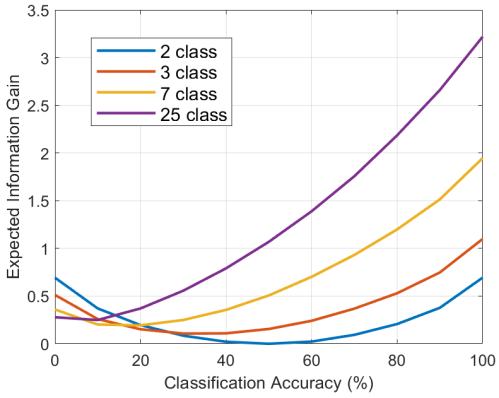

Figure 2: Expected information gain as a function of classification accuracy is shown for 2, 3, 7 and 25 classes from randomly generated confusion matrices.

This theorem entails that, for any given c, in the accuracy interval $[a, 1]$, expected information gain increases monotonically with classification accuracy. Hence, information gain is probabilistically a sound measure of classification accuracy in $[a, 1]$ where a decreases as c increases.

We experiment with four methods to fuse the posterior probabilities obtained from each classifier at the decision level: product, average, Bayesian, and majority voting. Let Z be the pattern to be assigned to one of the m classes after fusion.

Definition 2.3. Product [6, 16, 24]. Assign $Z \to \omega_j$ if $\prod_{i=1}^{n} P(\omega_j|x_i) = \max_{k=1}^{m} \prod_{i=1}^{n} P(\omega_k|x_i)$ [24]. The product rule for fusion assumes that the joint probability distribution of the measurements obtained from the classifiers are conditionally independent.

Definition 2.4. Average [6, 16, 24, 38]. Assign $Z \to \omega_j$ if $\frac{1}{n} \sum_{i=1}^{n} P(\omega_j|x_i) = \max_{k=1}^{m} \frac{1}{n} \sum_{i=1}^{n} P(\omega_k|x_i)$ [24]. The average rule for fusion assumes that the prior is equal.

Definition 2.5. Bayesian [19]. Assign $Z \to \omega_j$ if $P(\omega_j|x_1, \ldots, x_n) \approx \max_{k=1}^{m} \sum_{i=1}^{n} \sum_{l=1}^{m} P(\omega_k|\tilde{\omega}_l, \lambda_i) P(\tilde{\omega}_l|\lambda_i, x_i) P(\lambda_i|x_i)$ [19], where $\tilde{\omega}_l$ denotes the predicted class. The probabilities, $P(\omega_k|\tilde{\omega}_l, \lambda_i)$ and $P(\lambda_i|x_i)$, can be approximated from the confusion matrix of λ_i.

Definition 2.6. Majority voting [24, 28]. Assign $Z \to \omega_j$ if $\sum_{i=1}^{n} \Delta_{ji} = \max_{k=1}^{m} \sum_{i=1}^{n} \Delta_{ki}$ [24]. The term $\sum_{i=1}^{n} \Delta_{ki}$ adds the votes for the class ω_k from the individual classifiers. In case of equal votes for multiple classes, the class with the highest posterior probability is selected.

3 EXPERIMENTAL SETUP

3.1 Experiments

3.1.1 *Classification*

To evaluate classification accuracy with increase in number of modalities through an exhaustive comparison, we construct a tree with n leaf nodes, each represents a set containing one modality, and the root node represents the set of n modalities. The tree has n levels. The i^{th} level has $\binom{n}{i}$ nodes, $i = 1, \ldots, n$. A non-leaf node representing a set Q containing q modalities has q child nodes,

Algorithm 1 Selecting modalities using information gain

Inputs: $P_1, \ldots, P_n$.
Output: Set s.
Initialize: $S \longleftarrow \{M_{best}\}, S' \longleftarrow \{M_1, M_2, \ldots, M_n\} - S, M_{best}$ is the modality yielding highest recognition accuracy.
Compute $G^S(1) \longleftarrow G(T', T)$ using Eq. 2, where T' are the predicted class labels from M_{best}.
for $i = 2$ *to* n **do**
  for $j = 1$ *to* $|S'|$ **do**
    Compute predicted class labels, $T' \longleftarrow f(\{P_{M_i} : \lambda_{M_i} \epsilon S \bigcup S'_j\})$ such that f represents any late fusion operation and S'_j denotes the j^{th} element in set S'.
    Compute information gain $G^{S'}(j) \longleftarrow G(T', T)$ using Eq. 2.
  end for
  $G^S(i) \longleftarrow \max G^{S'}$
  $k \longleftarrow \arg_j \max G^{S'}$
  $S \longleftarrow S \bigcup S'_k$
  $S' \longleftarrow S' - S'_k$
end for
$k \longleftarrow \arg_i \max G^S$
Return $s \longleftarrow S_{1:k}$.

each representing a set $Q'_j \subset Q$ $(j = 1, \ldots, q)$ containing $q - 1$ modalities. Therefore, the total number of children of all nodes is: $\#Cases = \sum_{i=1}^{n-1} \binom{n}{i}(n - i)$.

3.1.2 Selection of modalities. The selection of modalities is carried out in three ways using information gain and correlation:
1) Use information gain, G, (Eq. 2) as a measure to compute the optimal combinations of $1, 2, \ldots, n$ modalities adding one at a time (ref. Algorithm 1). This is a greedy approach where the combination with the highest value of G is always selected.
2) Use correlation degree, ρ, [15, 34] as a measure to compute the optimal combinations of $1, 2, \ldots, n$ modalities adding one at a time. Select the subset with lowest ρ, similar to the approach in [15].
3) Assign a score, G_s, equal to its information gain (Eq. 2) to each modality and select the modalities with score greater than an appropriate threshold (0.2) which is determined experimentally and applies to all the datasets.

3.2 Training procedure

A total of 33 multilayer perceptron (MLP) classifiers (7×2 for DEAP, 5×2 for HCI Tagging, 3 for PAMAP2, 3 for UTD MHAD and 3 for Berkeley MHAD) are trained, one for each modality for each dataset. The hyperparameters (batchsize, number of layers, activation functions (tanh, relu), number of neurons in each layer, learning rate, dropout (for MLPs with more than one hidden layer)) for each MLP are fixed experimentally. We use a softmax classifier at the final layer and binary cross entropy as the cost function during training.

DEAP [25]: As in [42], the individual rated scales (1-9) are mapped to two levels of each valence and arousal states such that 1-5 is mapped to low and 5-9 to high emotion labels. The valence and arousal classification is done separately. As in [25, 42], the train/test set split is obtained by leaving one subject out cross-validation such that data from 31 subjects constitute the train set and data from the remaining subject constitute the test set. The mean classification accuracy from 32-fold cross-validation is reported. The preprocessed dataset available in MATLAB format in the downloaded dataset is used in our experiments. We extract six statistical features for all the 7 modalities considered in our experiments: means and standard deviations of the raw signals, means of absolute values of the first and second differences of the raw signals, and means of absolute values of the first and second differences of the normalized signals [37], from a 6-second window without overlap [42] such that each window constitutes a datapoint.

HCI Tagging [41]: The individual rated scales (1-9) are mapped to three levels of each valence and arousal states such that 1-3 is mapped to low, 4-6 to medium and 7-9 to high, as in [44]. The valence and arousal classification is performed separately. Two-third of the dataset is used for training, as in [43]. The experiments are repeated 10 times. The mean classification accuracy is reported.We extract 11 statistical features for all the 5 modalities considered in our experiments: the six features as in the DEAP dataset, skewness, kurtosis, min, max, and median [37, 42] from 6-second windows without overlap [42] such that each window constitutes a datapoint.

PAMAP2 [40]: We consider 12 actions for recognition, as in [40]. The train/test set split is obtained by leaving one subject out cross-validation; data from 8 subjects constitute the train set, as in [40]. The mean classification accuracy from 9-fold cross-validation is reported. We extract features from a 2-second window without overlap. Six statistical features as in the DEAP dataset are extracted from sensors 1 and 2. Three features are extracted from sensor 3 as in [9]: mean, variance, and standard deviation. The features are extracted for each dimension of the inertial signals and then concatenated.

UTD MHAD [9]: There are 27 action classes. Data from subjects $\{1, 3, 5, 7\}$ is used for training and subjects $\{2, 4, 6, 8\}$ for testing, as in [9]. The experiments are repeated 10 times and the mean of the classification accuracy over all the experiments is reported. We extract 8 statistical features from each 6-second window, as in [9], for inertial data across all the dimensions: mean, median, max, min, standard deviation, variance, skewness, kurtosis. Variance is extracted from 8 windows, as in [9], for each video for each dimension of the motion capture data. The features are computed over all dimensions, then concatenated over all the dimensions and windows. Depth motion map (DMM) features, as in [9] are extracted from the depth videos.

Berkeley MHAD [35]: There are 11 action classes. Data from the first 7 subjects are used for training and the last 5 for testing, as in [11, 35]. The experiments are repeated 10 times and mean classification accuracy is reported. Each inertial sequence is divided into 30 windows and each accelerometer sequence into 60 windows, as in [8]. Variance is extracted for inertial and accelerometer data from each window and across all the dimensions which is then concatenated over all the dimensions and windows. As in [8], DMM features constitute our feature vector for the depth videos.

4 EXPERIMENTAL RESULTS

Classification accuracy with respect to number of modalities. Our results show that an increase in the number of modalities

Table 1: Classification accuracy (%) reported for the best subset of modalities (i.e. the combination of modalities that yields highest accuracy from all fusion methods) for each of the three modality selection methods (G, ρ, G_s). The "Exhaustive" column reports the highest classification accuracy obtained by considering all possible combinations of modalities (total # combinations = $\sum_{i=1}^{n} \binom{n}{i}$). All modalities present in a dataset are mentioned below the dataset name. The fusion method and the subset of modalities yielding the highest accuracy are reported. The baselines for similar experimental conditions as ours are shown. The highest accuracy for each dataset is highlighted.

Dataset	Exhaustive	G	ρ	G_s	Baseline
DEAP (valence) eeg, gsr, resp, temp, plet, emg, eog	**75.72** - Product eeg, eog, emg, resp	**75.72** - Product eeg, eog, emg, resp	75.41 - Bayesian eeg, eog, emg, plet	**75.72** - Product eeg, eog, emg, resp	57.6 [25]
DEAP (arousal) eeg, gsr, resp, temp, plet, emg, eog	**66.41** - Bayesian eeg, eog, emg, gsr, resp	66.40 - Bayesian eeg, eog, gsr	65.30 - Bayesian eeg, eog, emg	66.31 - Bayesian eeg, eog	62 [25]
HCI Tagging (valence) eeg, ecg, gsr, resp, temp	**68.63** - Product eeg, ecg, gsr, resp	**68.63** - Product eeg, ecg, gsr, resp	**68.63** - Product eeg, ecg, gsr, resp	**68.63** - Product eeg, ecg, gsr, resp	56.83 [44]
HCI Tagging (arousal) eeg, ecg, gsr, resp, temp	**66.36** - Average eeg, ecg, gsr, resp	**66.36** - Average eeg, ecg, gsr, resp	66.19 - Average eeg, ecg, gsr, temp	**66.36** - Average eeg, ecg, gsr, resp	54.73 [44]
PAMAP2 s1, s2, s3	**90.98** - Product s1, s2, s3	83.57 - Product s1, s2	**90.98** - Product s1, s2, s3	**90.98** - Product s1, s2, s3	89.24 [39]
UTD-MHAD depth, skel, iner	**92.40** - Product depth, skel, iner	**92.40** - Product depth, skel, iner	78.39 - Product skel, iner	**92.40** - Product depth, skel, iner	79.1 [9]
Berkeley MHAD depth, skel, iner	97.05 - Average depth, skel, iner	97.05 - Average depth, skel, iner	94.22 - Product depth, skel	97.05 - Average depth, skel, iner	**98.23** [8]

Table 2: Percentage of *#Cases* (ref. Section 3.1.1) where classification accuracy decreases as a modality is added. The range of accuracy (%), stated within parentheses, is obtained from all possible combination of modalities (# combinations = $\sum_{i=1}^{n} \binom{n}{i}$).

Method	DEAP (valence)	DEAP (arousal)	HCI Tagging (valence)	HCI Tagging (arousal)	PAMAP2	UTD -MHAD	Berkeley MHAD
Product	47.85 (15.68)	64.17 (8.35)	18.67 (25.83)	16 (23.56)	0 (20.26)	0 (29.19)	0 (16.66)
Average	47.85 (15.60)	64.40 (8.35)	18.67 (25.44)	14.67 (23.72)	33.33 (13.32)	0 (24.08)	33.33 (18.91)
Bayesian	34.01 (15.66)	48.30 (8.66)	6.67 (22.62)	6.67 (20.62)	22.22 (11.69)	0 (23.87)	22.22 (17.57)
Voting	47.39 (15.15)	49.21 (8.41)	46.67 (19.9)	41.33 (20.15)	0 (9.16)	0 (22.50)	0 (14.19)

may not increase the classification accuracy, especially for emotion recognition from physiological signals (ref. Table 1 and Table 2). Adding more modalities for fusion might add noise and create confusion leading to misclassification. The decrease in classification accuracy with increasing modalities is observed for all the fusion methods and all datasets except UTD-MHAD dataset, as shown in Table 2. Of the four fusion methods, Bayesian fusion yields the smallest decrease in accuracy with increase in number of modalities compared to other fusion methods for all the datasets. This is because Bayesian fusion takes into account the uncertainty of the classifiers for each class by combining the classifiers as a weighted combination of the error distribution over the classes.

Evaluation of different metrics for selecting modalities. Our results (ref. Table 1) show that, for all the datasets, the classification accuracy obtained from the subset of modalities selected using information gain is closer to the highest accuracy than that selected using correlation degree. Hence, information gain outperforms correlation degree as a criterion for modality selection. The lowest absolute difference between the true highest accuracy ("Exhaustive") and the accuracy obtained using selected modalities, added over all datasets, is 1.79 for product fusion, 1.7 for average fusion, 0.03 for Bayesian fusion, and 0.01 for majority voting, obtained using metrics G_s, G, G and G respectively. Since information gain increases monotonically with classification accuracy (ref. Theorem 1), it is a useful metric for selecting a subset of modalities that will yield high accuracy. As shown in Table 1, the subset of modalities selected using different metrics always yields a classification accuracy comparable, if not equal, to the highest accuracy.

The correlation degree criteria initially selects the modality highly correlated with the true class labels, then it selects modalities least correlated with the selected modalities. This helps in reducing redundancy. However, it can reduce relevant information as well which can lower the classification accuracy, as observed from our results. On the other hand, selecting modalities based on their individual score using information gain and filtering them using a threshold, allows selection of modalities highly correlated with the true class labels. This preserves relevant information but might have high redundancy. However, it outperforms correlation degree as a selection criteria, as seen from our experiments.

Information gain modality selection method (ref. Algorithm 1) selects the combination of modalities after fusion that has the highest correlation with the true class label. This yields the highest classification accuracy, comparable with the true best combination in our experiments. Algorithm 1 requires fusing the modalities

before computing the information gain. Hence, it depends on the fusion method while the other two, correlation degree and filtering using information gain, are independent of the fusion method.

5 CONCLUSIONS

In this paper, we investigated the optimal modality selection problem for time-series data in the context of late fusion. We analyzed multimodal emotion or action classification using four late fusion methods and five benchmark datasets. Our experimental analysis on product, average, Bayesian and majority voting late fusion methods show that the fusion methods perform differently based on the posterior distribution estimated by each modality. Our results show that for different fusion methods, increasing the number of modalities might not necessarily increase the classification accuracy. We analyze multiple methods for selecting a subset of modalities for late fusion and observe that information gain is an useful measure for selecting modalities which is consistent for all the datasets. The classification accuracy obtained from the selected subset of modalities is comparable to the highest accuracy in all cases.

REFERENCES

[1] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16, 6 (2010), 345–379.

[2] Pradeep K Atrey, Mohan S Kankanhalli, and John B Oommen. 2007. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 3, 1 (2007), 2.

[3] B. Banerjee and J. K. Dutta. 2014. SELP: A general-purpose framework for learning the norms from saliencies in spatiotemporal data. *Neurocomput.* 138 (2014), 41–60.

[4] M. Baruah and B. Banerjee. 2020. A multimodal predictive agent model for human interaction generation. In *CVPR Workshop*.

[5] M. Baruah and B. Banerjee. 2020. The perception-action loop in a predictive agent. In *CogSci*.

[6] C. Busso, Z. Deng, et al. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. 6th international conference on Multimodal interfaces*. ACM, 205–211.

[7] Ginevra Castellano, Loic Kessous, and George Caridakis. 2008. Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction*. Springer, 92–103.

[8] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Transactions on Human-Machine Systems* 45, 1 (2015), 51–61.

[9] C. Chen, R. Jafari, and N. Kehtarnavaz. 2015. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *IEEE International Conference on Image Processing*. 168–172.

[10] Alok Kumar Chowdhury, Dian Tjondronegoro, Vinod Chandran, and Stewart G Trost. 2018. Physical Activity Recognition Using Posterior-Adapted Class-Based Fusion of Multiaccelerometer Data. *IEEE journal of biomedical and health informatics* 22, 3 (2018), 678–685.

[11] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.

[12] J. K. Dutta and B. Banerjee. 2014. Learning features and their transformations from natural videos. In *IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments*. Orlando, FL, 55–61.

[13] J. K. Dutta and B. Banerjee. 2015. Online detection of abnormal events using incremental coding length. In *AAAI*. 3755–3761.

[14] J. K. Dutta, B. Banerjee, and C. K. Reddy. 2016. RODS: Rarity based outlier detection in a sparse coding framework. *IEEE Trans. Knowl. Data Eng.* 28, 2 (2016), 483–495.

[15] K. Goebel, W. Yan, and W. Cheetham. 2002. A method to calculate classifier correlation for decision fusion. *Proceedings of decision and control* (2002), 135–140.

[16] Hatice Gunes and Massimo Piccardi. 2005. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, Vol. 4. IEEE, 3437–3443.

[17] D. Hu, C. Wang, F. Nie, and X. Li. 2019. Dense multimodal fusion for hierarchically joint representation. In *ICASSP*. 3941–3945.

[18] Volkan Isler and Ruzena Bajcsy. 2005. The sensor selection problem for bounded uncertainty sensing models. In *Proceedings of the 4th international symposium on Information processing in sensor networks*. IEEE Press, 20.

[19] Yuri Ivanov, Thomas Serre, and Jacob Bouvrie. 2005. Error weighted classifier combination for multi-modal human identification. (2005).

[20] Mohan S Kankanhalli, Jun Wang, and Ramesh Jain. 2006. Experiential sampling on multiple data streams. *IEEE transactions on multimedia* 8, 5 (2006), 947–955.

[21] M. H. Kapourchali and B. Banerjee. 2018. Unsupervised feature learning from time-series data using linear models. *IEEE Internet of Things Journal* 5, 5 (2018), 3918–3926.

[22] M. H. Kapourchali and B. Banerjee. 2019. State estimation via communication for monitoring. *IEEE Trans. Emerg. Topics Comput. Intell.* (2019).

[23] M. H. Kapourchali and B. Banerjee. 2020. EPOC: Efficient perception via optimal communication. In *AAAI*.

[24] J. Kittler, M. Hatef, R. Duin, and J. Matas. 1998. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence* 20, 3 (1998), 226–239.

[25] S. Koelstra, C. Muhl, et al. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.

[26] Ludmila I Kuncheva, Christopher J Whitaker, Catherine A Shipp, and Robert PW Duin. 2000. Is independence good for combining classifiers?. In *icpr*. IEEE, 2168.

[27] K. Lam, R. Cheng, B. Liang, and J. Chau. 2004. Sensor node selection for execution of continuous probabilistic queries in wireless sensor networks. In *Proc. ACM 2nd international workshop on Video surveillance & sensor networks*. 63–71.

[28] Utthara Gosa Mangai, Suranjana Samanta, Sukhendu Das, and Pinaki Roy Chowdhury. 2010. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical review* 27, 4 (2010), 293–307.

[29] T. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

[30] S. Najnin and B. Banerjee. 2015. Improved speech inversion using general regression neural network. *The Journal of the Acoustical Society of America* 138, 3 (2015), EL229–EL235.

[31] S. Najnin and B. Banerjee. 2017. A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Commun.* 92 (2017), 24–41.

[32] S. Najnin and B. Banerjee. 2019. Speech recognition using cepstral articulatory features. *Speech Commun.* 107 (2019), 26–37.

[33] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[34] Gang Niu, Tian Han, Bo-Suk Yang, and Andy Chit Chiow Tan. 2007. Multi-agent decision fusion for motor fault diagnosis. *Mechanical Systems and Signal Processing* 21, 3 (2007), 1285–1299.

[35] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 53–60.

[36] P. Pahalawatta, T. Pappas, and A. Katsaggelos. 2004. Optimal sensor selection for video-based target tracking in a wireless sensor network. In *International Conference on Image Processing*, Vol. 5. IEEE, 3073–3076.

[37] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[38] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[39] Attila Reiss and Didier Stricker. 2012. Creating and benchmarking a new dataset for physical activity monitoring. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 40.

[40] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium on*. IEEE, 108–109.

[41] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55.

[42] Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshu Mittal, and Samit Bhattacharya. 2017. Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset.. In *AAAI*. 4746–4752.

[43] Mimoun Ben Henia Wiem and Zied Lachiri. 2016. Emotion assessing using valence-arousal evaluation based on peripheral physiological signals and support vector machine. In *Control Engineering & Information Technology (CEIT), 2016 4th International Conference on*. IEEE, 1–5.

[44] Mimoun Ben Henia Wiem and Zied Lachiri. 2017. Emotion Classification in Arousal Valence Model using MAHNOB-HCI Database. *Int. J. Adv. Comput. Sci. Appl. IJACSA* 8, 3 (2017).

[45] Yi Wu, Edward Y Chang, Kevin Chen-Chuan Chang, and John R Smith. 2004. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 572–579.