Sparse F-IncSFA for Action Recognition

Chukiong LOO* and Yousefi BARDIA**
Department of Artificial Intelligence
University of Malaya
Kuala Lumpur, Malaysia
ckloo.um@um.edu.my*

High dimensional input streams and unsupervised learning are two important factors in the area of humanoids and processes of the actions and movements of human. Our Fast Incremental Slow Feature Analysis (F-IncSFA) can learn and extract the few significant features of the complex sensory input sequences regarding high-level spatio-temporal conceptions. In this paper, the application of the F-IncSFA and some of its structure to make a hierarchical compound network made of F-IncSFA has been described. Also the techniques developed by adding efficient sparse coding as an encoder and a preprocessing step before an application of the F-IncSFA. The efficient sparse coding can dramatically reduces the dimension of extracted features and outcome of the efficient sparse coding are quite small as compared with the size of high-dimension video obtained by humanoid or human action. It has revealed excellent and promising dimension reduction by this preprocessor.

***Key Words*:** Sparse fast incremental slow feature analysis (Sparse-F-IncSFA), unsupervised learning, hierarchical network, efficient sparse coding.

1. Introduction

The ability of human beings for getting several proficiencies derived from communication through the surroundings even with no interference as a teacher. Here for dimension reduction of video input, the novel unsupervised learning system is proposed which called sparse-F-IncSFA that it is a combination of sparse coding[12] plus Fast Incremental Slow Feature Analysis(F-IncSFA)[14] for spatiotemporal features extraction. The sparse coding carries out compressing whereas F-IncSFA utilized for spatiotemporal feature extraction which modify gradually over time. Legenstein et al. [1] have revealed a alike two phase learning system created of a hierarchical slow feature analysis (H-SFA) network [2]. As the batch method is not suitable for developmental learning though, the presented approach as incremental technique for learning the behavior of the high dimensional input streams.

The rest of this paper is organized as follows. This section was introduction, a review SFA, and sparse coding will be present at next subsections. Sections 2 introduce our method structure of F-IncSFA and sparse coding. Section 3, 4 are results and conclusion, respectively.

1.1 Slow Feature Analysis (SFA)

Slow Feature Analysis (SFA) is one of the unsupervised learning methods. The functions which planning the input stream to the most slowly changing outcomes are characteristic of a number of elementary representatives of world possessions, summarizing away unrelated details selected up by the sensors [3]. Moreover, considering a mobile agent which has high-dimensional video input can be a searching an otherwise stationary room and encode the data by using the combining the situation and direction by slow features [4].

SFA usually concerns with the optimization complexity: as it is common that for identification of **x(t)** as input by **D** dimension, $x(t)=[x_1(t),...,x_D(t)]^T$, there is a set of functions like f(x) which has L dimension, $f(x)=[f_1(x),...,f_L(x)]^T$, which can produce the output by **L** dimension as **y(t)** like, $y(t)=[y_1(t),...,y_L(t)]^T$. So the relation between these set is $y_l(t) := f_l(x(t))$.

$$\Delta_l := \Delta(y_l) := \left\langle \dot{y}_l^2 \right\rangle \text{ is minimal} \quad (1)$$

$$\langle y_l \rangle = 0 \text{ (Zero mean)}, \quad (2)$$

$$\left\langle y_l^2 \right\rangle = 1 \text{ (Unit variance)}, \quad (3)$$

$$\forall d < l : \langle y_d y_l \rangle = 0 \text{ (De-correlation and order)}, \quad (4)$$

These general definitions as 2, 3 are the restrictions for having insignificant constants in output and 4 is for de-correlation restrictions for features which are same do not be coded, respectively. By means of $\dot{y}$ and<·>is a representation of evaluation for derivative of y and the average of sequential, correspondingly. The problem will be defined by finding the $f(x)$ for generating the slow varying output.

It is noticeable that for solution of this problem the optimization of variation calculus like [5] is not applicable but it is mostly straightforward especially for eigenvector method. By considering f_l is constrained to be a linear function which made of combination of a finite set nonlinear functions p so for output function we will have:

$$y_l(t) = f_l(x(t)) = w_d^T\, p(x(t)) \quad (5)$$

Then we will have $z(t) = p(x(t))$ .By the changes which previously done, the optimization problem will be introduced by minimizing (6) by finding the w_l.

$$\Delta(y_l) = \left\langle \dot{y}_l^2 \right\rangle = w_l^T \left\langle \dot{z}\dot{z}^T \right\rangle w_l \quad (6)$$

If the p functions are selected such that z has unit of covariance matrix and zero mean, the three restrictions will be satisfied if and only if the weight vectors have orthonormal difference. The whitening is the very common technique which is used for finding the p. For whitening the principle component of the input data is required that by

This project is sponsored by Flagship grant from University of Malaya (FL0016-2011))

considering the zero mean and individuality covariance matrix put the **x** to **z** and by this **z** the SFA problem will be converted to the linear problem. By considering the equation (6) for minimizing the **L** norm set of eigenvectors of $\langle \dot{z}\dot{z}^T \rangle$. The desired features will be obtained from a set of principle components of $\dot{z}$.

Specified an input by means of two parts that differ rapidly over time (e.g., **x(t)** given by $\breve{x}_1(t) = \sin(t) + \cos(11t)^2$, $\breve{x}_2(t) = \cos(11t)$. $t \in [0, 2\pi]$), SFA will find the slowest feature hidden in the signal (here: $y_1(t) = \breve{x}_1(t) - \breve{x}_2(t) = \sin(t)$). Occasionally, the slowest component is not the most spontaneous one; for instance while examination of a thing which has a motion watched by a camera and irregularly disappears the field observation; the slowest feature is the existence/nonexistence of the object, not its situation.

Figure 1[6, 14] shows a structure of Hierarchical network of F-IncSFA on a straightforward reproduced model and interaction a video streams. Hierarchical Fast IncSFA (H-F-IncSFA) extracts a slow feature which codes directly for the point of the interaction. In [6, 14] a performance of H-F-IncSFA to process data comprises of high dimensional image considered. H-F-IncSFA has 17 units extend in three layers, every layer taught in sequence beginning bottom to up. H-F-IncSFA does not require for collect and set aside for covariance matrix or input data and is consequently appropriate to open-ended learning development.

On the other hand, H-IncSFA distillery does not solve the problems relating to the consequence of spatially in important and gradually changing external parts. Additionally, the elevated layers of the H-IncSFA system require the subordinate layers to converge in initial. Therefore additional models are involved by the network is entirely well-designed.

1.2 Efficient sparse coding algorithms

Sparse coding gives a class of algorithms for encountering stated briefly but clearly demonstrations of stimuli; specified merely un-labeled data from input, the fundamental functions which get the features from the data by having higher level can be learned by it. While the natural images are the application of sparse coding algorithm, the accessible fields of neurons in the visual cortex will be the base of learning[7, 8]; furthermore, for the video sparse coding can make localized bases [9]. Contrasting several additional unsupervised learning methods like PCA, sparse coding is able to be used for learning the sets which are over-completed; in the input dimension has fewer amounts of bases.

Although the high intention of declaration sparse coding models, we take to be true that their improvement has been disadvantaged by their expensive computational cost. In detail, learning big, highly over-complete showing has been tremendously expensive. Here, a class of efficient sparse coding algorithms which are based on alternating optimization over two subsets of the variables is addressed. In this case, the problem of optimization over the each group

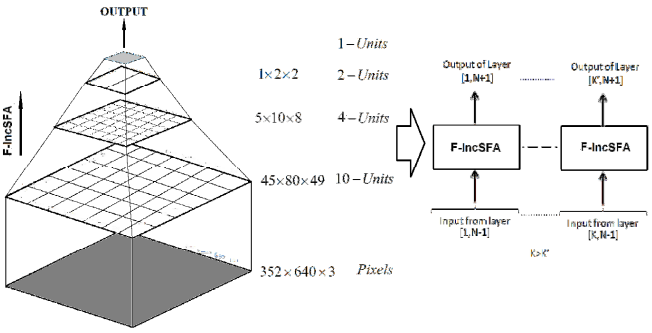

Figure 1. The figure depicts the Hierarchical Network designed based on Fast Incremental Slow Feature Analysis (F-IncSFA).

of video set depends on the time. The optimization will be utilized for dimension reduction of the input video.

1.3 Learning bases applying the Lagrange dual

In this part, for our feature reduction we use the problem like optimization problem over bases B specified fixed coefficients S. This lessens to the subsequent problem:

$$\text{Minimize } \|X - BS\|_F^2 \qquad (7)$$

$$\sum_{i=1}^{k} B_{i,j}^2 \le c, \quad \forall j = 1,...,n \qquad (8)$$

This is a smallest amount of squares problem with quadratic constraints. Totally, this hampered optimization problem can be solved applying gradient descent through iterative projection [10]. On the other hand, by applying a Lagrange dual it will be better solved and by applying conjugate gradient or Newton's technique the Lagrange dual can be optimized [14]. The optimal bases B will be attained as follows:

$$B^T = (SS^T + \Lambda)^{-1}(XS^T) \qquad (9)$$

Where $\Lambda = diag(\vec{\lambda})$ and each $\lambda_j \ge 0$ is a dual variable. The benefit of the dual solution is applying it in smaller number optimization variables as compare with the original [14].

Totally, A sparse coder gets an input $x \in R^d$ and map s it to the latent representation $h \in R^d$, B obtained by sparse coding and the **h(t)** is the sparse coding function for mapping the input **x(t)** to the matrix B.

2. Method (Sparse F-IncSFA)

While sparse coding can shortly present of the correlated parts of input, and F-IncSFA is able to take out significant variant in time, the proposed approach presents for reduction of dimension, space and time, for a robot's visual input:

1) Input Signal: obtain the present raw I-dimensional input like vector x(t).

2) Normalization: For normalization of the input to attain

$$x(t) = [x_1(t),...., x_l(t)] \qquad (10)$$

$$x_i(t) := \frac{x_i(t) - < \breve{x}_i >}{F} \qquad (11)$$

Where, **F** is an upper bound of **x**

$$\text{So that} \quad < \breve{x}_i >= 0 \qquad (12)$$

$$\text{and} \quad 0 \le \breve{x}_i < 1 \qquad (13)$$

3) Sparse Coding Update: For every pattern of input as **x(t)**, suppose the reconstruction B and remaking the model of weights. The weights are applied to obtain the code **h(t)**.

4) Updating of F-IncSFA:

a) Whitening by CCIPCA: The hidden unit activations **h(t)** are normalized to produce **z(t)** with zero mean and character covariance matrix **I**. This so-called whitening is able to be done incrementally by using Candid Covariance-free Incremental Principal Component Analysis (CCIPCA) on **h(t).**

b) Derivative signal: To attain the difference of the signal over time, **z(t)** is distinguished by respect to **t** to produce $\dot{z}(t)$. We apply the variation over a single time step as a fast estimate of the derivative.

c) Slow Features: By applying incremental minor component analysis to the matrix $< z\dot{z}^T >$**, J** eigenvectors by the lowest eigen-values λ_i are extracted. These are the recent evaluates of the slow features; **W(t).**

5) Output: Then, **y(t) = z(t) W(t)** is the Sparse-F-IncSFA output.

3. Results

3.1 Humanoid Action Viedo Stream

For ability to adjust configuration and size to fit new conditions of F-IncSFA analysis, the high-dimensional video stream has been performed by some images from our datasets [3].

3.2 Hierarchical Fast IncSFA(H-F-IncSFA)

We want to calculate features that code from different expressions. This can be an inspiring yet difficult to handle due to the nonlinear function for mapping the different poses and structures. The proposed approach applies a hierarchical model which motivated from human visual system like Franzius et al. work [4]. The structure and architecture of hierarchical model has shown in the figure 1 (a) prepared in some layers of multiple F-IncSFA units, smaller dimensions by overlapping of different fields.

On the lowest layer, the accessible every module area comprises of an image of 352 × 640 pixels. Then, the production from the initial level shapes 45 × 80 × 50 grids for each module 10 slow features. By extending over interested areas of 5 × 10 × 8, this layer's output, with 4 slow features per module, becomes a 1 × 2 × 2 grid. Correspondingly, the layers third and fourth reduce the dimension of image to create one very small output. The network sequentially trains from bottom to up over the whole dataset. Usually, the amount of successfully extraction of slow feature is merely restricted by the presentation of CCIPCA as compared to batch PCA. The batch PCA is able to potentially improved extraction of the eigenvectors related the smallest eigenvalue because it can iterate over all of the information. On the other hand, this data, while approved on to our second part, is hardly suitable (for the input information, small eigenvalue ways can normally be eliminated).

Figure 2 illustrates the some application results our proposed approach (F-IncSFA). The outcome of applying F-IncSFA to the video sequence in the shape of H-F-IncSFA which includes some images have been obtained and presented. The figure 2 and figure 3 represents two different structure of applying F-IncSFA. The output has been shown in the various structures for both of actions.

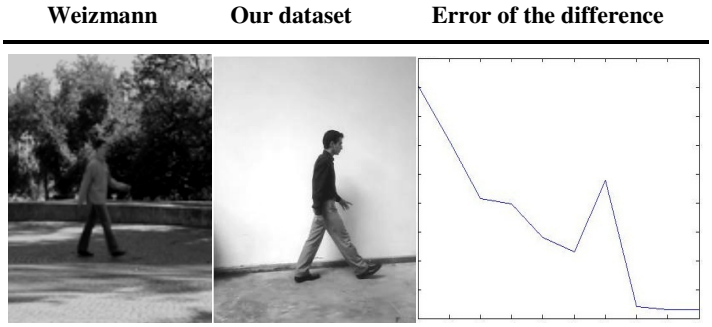

Figure 2. Experimental of the H-F-IncSFA technique is shown; (a) the images as a sample of the Weizmann Action dataset and (b) presents our testing dataset. By applying the H-F-IncSFA the error graph will be obtained from output layer of the network which is shown in (c).

3.3 Human Interaction Experiment

For evaluation of the performance of Sparse coding and the state-of-the art batch H-SFA network which previously presented and as applications of the two methods, here, applying these two methods as results of application of the human video streams. The H-F-IncSFA has been applied to Weizmann Action dataset; a sample of attained results has been presented in the figure 2. The figure 2(a) and (b) show the samples of two datasets for applying the hierarchical network. The figure 2(c) reveals the mean square error of between the outputs of these two tests.

To evaluate the application of both techniques, the results of the Sparse coding is still required. The figure 3 presents the outcome applying the Sparse coding to the walking position for Weizmann Action dataset and our datasets. By applying the sparse coding to the dataset the Matrix B will be obtained which will be utilized for the F-IncSFA section. The figure 3(a) is revealed the application of the sparse coding to the Weizmann Action dataset. As it is presented in the Figure 3

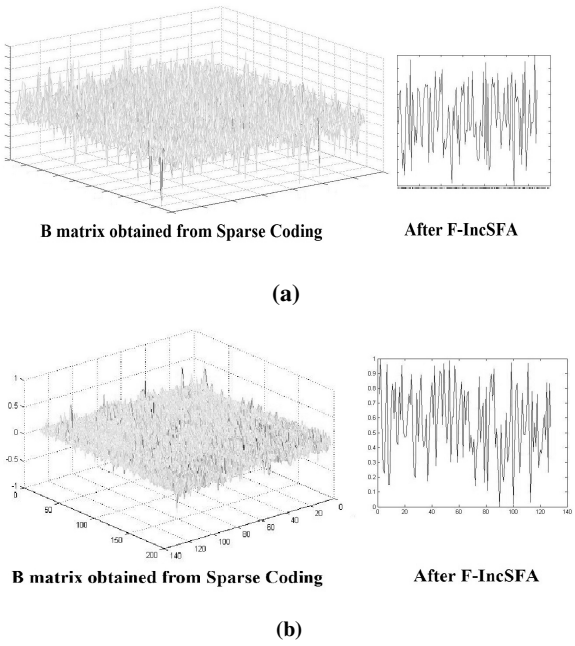

Figure3. The figure represents the outcome of the efficient sparse coding for two different datasets. (a) Shows the results of applying the Sparse to the Weizmann Action dataset (Walking); (b) reveals this application for our datasets.

regarding the applications of H-F-IncSFA and Sparse coding to the two different dataset and the Weizmann Action dataset and the figure 3(b) represents its usage for mentioned datasets.

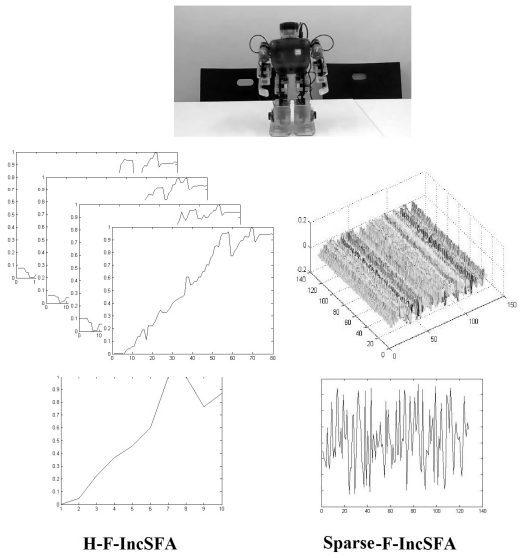

Figure4. The figure shows the outcome of the efficient sparse coding and Sparse-F-IncSFA for high dimension video stream dataset from hominid robot.

3.4 Objects Interaction Experiment

For ability to adjust configuration and size to fit new conditions of F-IncSFA analysis, the high-dimensional video stream of the human interaction has been performed by last subsection. Previous section presents a sample image attained from our datasets. One significant area of applying the sparse coding is considering the application for object interaction. For this aim, the sparse coding has been applied to the other dataset obtained from the robot interactions. Two H-F-IncSFA and sparse coding have applied to our robot video streams. The initial size of the video streams for entering to our H-F-IncSFA and sparse coding is 89 Mb that it will be dramatically diminished after sparse coding and the attained B matrix which is as input of F-IncSFA will has 188Kb size whereas, there is no dimension reduction in H-F-IncSFA method.

4. Conclusion

The proposed application of the novel unsupervised learning method F-IncSFA was the moderation of the latest online algorithm for Slow Feature Analysis [13, 14] and the hybrid form of this technique in the shapes of H-F-IncSFA and sparse coding has been presented. High dimensional input streams as one of the important factors in the area of humanoids and processes of the actions and movements of human have been analyzed by applying the compound configurations of our F-IncSFA. Furthermore results of development of F-IncSFA by adding efficient sparse coding as a preprocessing step have been attained. The efficient sparse coding dramatically reduced the dimension of features set and it is proved that the outcome of the efficient sparse coding are quite small as compared with the size of high-dimension video obtained by humanoid or human action plus it can be a good start for the view independent motion action recognition.

References

[1] Legenstein, R., Wilbert, N. and Wiskott, L. “Reinforcement learning on slow features of high-dimensional input streams,” *PLoS Computational Biology,* vol. 6, no. 8, p. e1000894, 2010.

[2] Franzius, M., Sprekeler, H. and Wiskott, L. “Slowness and sparseness lead to place, head-direction, and spatial-view cells,” *PLoS Computational Biology*, vol. 3, no. 8, p. e166, 2007.

[3] Wiskott, L., Zito, T., Wilbert, N. and Berkes, P. “Modular toolkit for data processing (mdp): a python data processing framework” *Frontiers in Neuroinformatics*, 2, 2008.

[4] Jolliffe, I. T. “Principal Component Analysis” *Springer-Verlag, New York*, 1986.

[5] Zhang, Y., Weng, J. “Convergence analysis of complementary candid incremental principal component analysis” *Michigan State University*, 2001.

[6] Kompella, V., Luciw, M. D. and Schmidhuber, J. “Incremental slow feature analysis,” in *IJCAI*, pp. 1354–1359, 2011.

[7] Olshausen, B. A. and Field, D. J. “Emergence of simple-cell receptive field properties by learning a sparse code for natural images” *Nature*, 381:607–609, 1996.

[8] Olshausen, B. A. and Field, D. J. “Sparse coding with an overcomplete basis set: A strategy employed by V1” *Vision Research*, 37:3311–3325, 1997.

[9] Lewicki, M. S. and Sejnowski, T. J. “Learning over-complete representations” *Neural Comp*., 12(2), 2000.

[10] Olshausen, B. A. “Sparse coding of time-varying natural images” *Vision of Vision,* 2(7):130, 2002.

[11] Censor, Y. and Zenios, S. A. “Parallel Optimization: Theory” *Algorithms and Applications*. 1997.

[12] Lee, H., Battle, A., Raina, R., Ng, A. Y. “Efficient sparse coding algorithms” *NIPS*, pp 1-8, 2006.

[13] Klemm, R. “Adaptive airborne mti: an auxiliary channel approach” *Proc Inst Elect Eng F*, vol.134, pp.269–276, 1987.

[14] Yousefi, B., Loo, C.K. “Development of Fast Incremental Slow Feature Analysis (F-IncSFA)” *International Joint Conference on Neural Networks, IJCNN 2012*. Unpublished.