in: Artifical Neural Networks - ICANN 09, Limasol, Cyprus, Part II, LNCS 5769, pp. 735-744, Springer 2009

Adaptive Feature Transformation for Image Data from Non-stationary Processes

Erik Schaffernicht[1], Volker Stephan[2], and Horst-Michael Gross[1]

[1] Ilmenau University of Technology
Neuroinformatics and Cognitive Robotics Lab
98693 Ilmenau, Germany
[2] Powitec Intelligent Technologies GmbH
45219 Essen-Kettwig, Germany
Erik.Schaffernicht@Tu-Ilmenau.de

Abstract. This paper introduces the application of the feature transformation approach proposed by Torkkola [1] to the domain of image processing. Thereto, we extended the approach and identifed its advantages and limitations.
We compare the results with more common transformation methods like Principal Component Analysis and Linear Discriminant Analysis for a function approximation task from the challenging domain of video-based combustion optimization. It is demonstrated that the proposed method generates superior results in very low dimensional subspaces.
Further, we investigate the usefulness of an adaptive variant of the introduced method in comparison to basic subspace transformations and discuss the results.

1 Introduction

Optimizing the combustion of coal in power plants is an important task, since increasing efficiency equals a reduction of carbon oxides (CO and CO2), nitrogen oxides (NOx) and other greenhouse gases in the flue gas. But all data normally measured at a plant is insufficient to build meaningful models and controllers. Therefore, our approach includes cameras to actively observe the flame itself. On one hand, with this additional information about the combustion process improved controllers can be built automatically. On the other hand, relying on image data introduces additional challenges.

The use of the original pixel space for learning algorithms that operate on image data is a rare occurrence. The high dimensionality of this space is a major obstacle in this respect, because this leads to a high complexity of the learning problem and a high number of free parameters to be estimated for an approximation or classification task. Furthermore, the feasibility of this approach is restricted by the computational effort required to handle the data.

Hence, preprocessing is applied to extract useful information from the original images. One way to achieve this is the use of designed features like certain geometric shapes, intensity values or certain texture patterns. This implicitly

requires at least a bit expert knowledge by the system designer to decide which methods are meaningful for the given problem.

Another way to cope with the problem are feature transformation algorithms which attempt to find an image subspace that contains much useful information. Typically these methods are guided by a statistical criterion to achieve this goal. Perhaps the best known representatives are *Principal Component Analysis* (PCA) [2], *Independent Component Analysis* (ICA) [3], *Nonnegative Matrix Factorization* (NMF) [4] and *Linear Discriminant Analysis* (LDA) [2]. The basic forms of these algorithms produce linear transformations only, but there are several non-linear (e.g. kernel-based) extensions for all methods, but PCA and ICA specifically attracted a lot of attention in this respect.

The PCA transforms data into a subspace based on the eigenvectors of the data covariance matrix, hence this produces axes along the most variant parts of the data. High eigenvalues mark high variant directions. The resulting subspaces are often named according to the task, like *eigenfaces* or, for combustion optimization, *eigenflames*. This technique, as well as ICA and NMF, are purely data driven. They only consider the data intrinsic relations, but not the recognition or approximation task to be solved. ICA tries to find subspaces that represent independent data parts. A contrast function like *Negentropy* or *Mutual Information* is used to measure the independence of the resulting subspace dimensions. The NMF transformation's unique selling point is that all subspace dimensions and resulting data points are in fact non negative, which is a constraint for certain application areas.

Unlike the aforementioned methods, algorithms like the LDA take the target of the learning problem into account to find a suitable subspace representation. It derives itself from the Fisher criterion [5] and aims at a subspace transformation that allows a good approximation with linear learning machines.

The *Maximal Mutual Information* (MMI) transformation introduced by Torkkola [1] is similar in this respect. It takes the target values into account, but unlike the LDA it does not make any assumptions about a specific learning machine. Instead, it tries to maximize the information content about the target in the new subspace. The basic ideas and mechanisms of this approach are recapped in Sect. 2.

The application of this approach to image data is straightforward, but requires the consideration of its limitations for this high dimensional domain. Additionally, we propose an supplemental step in the algorithm to capture image specific traits. A comparison to PCA approaches on a flame image prediction task completes Sect. 3.

Since our intended application area, the intelligent control of combustion processes in power plants, is non-stationary, the feature extraction's requirements include a certain adaptivity. A comparison of different initializations, PCA and LDA is given, and we will discuss the use of the MMI transformation as adaptive system and the pitfalls associated in Sect. 4.

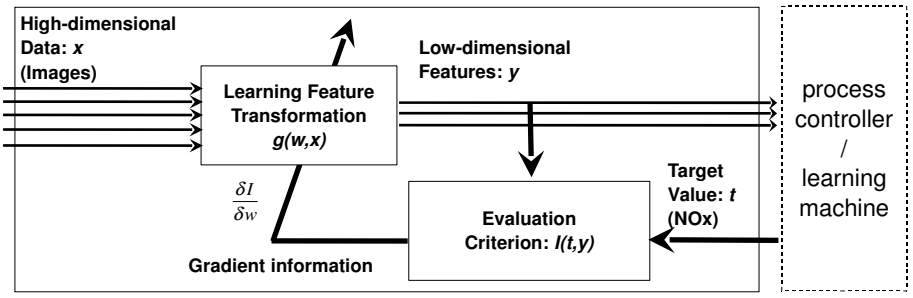

Fig. 1. The original image data x is transformed by some function g into a lower-dimensional space. An evaluation criterion, the Quadratic Mutual Information, measures the correspondence to the desired target value t, e.g. the nitro oxides to the reduced images y. From this criterion, a gradient information $\delta I/\delta w$ is derived and used to adapt the transformation parameters w.

2 Feature Extraction Using Mutual Information Maximization

The Maximal Mutual Information approach of Torkkola [1] is built upon the *Information-Theoretic Learning* (ITL) framework introduced by Principe [6]. The idea is to find a transformation that maximizes the *Mutual Information* (MI) between the transformed data in a certain subspace and the desired target values. A number of "forces" is computed to be used as the direction in a gradient ascent to maximize the MI.

The basic adaption loop for the optimization process is shown in Fig. 1. The original input data sample x_i is transformed by some transformation g with the parameters w into a lower dimensional space. The transformed data is denoted by y_i. The goal is to find those transformation parameters w that confer the most information into the lower dimensional space with respect to the target.

The update rule for the parameters of the transformation is given by the following equation, where α denotes the learning rate

$$w_{t+1} = w_t + \alpha \frac{\partial I}{\partial w} = w_t + \alpha \sum_{i=1}^{N} \frac{\partial I}{\partial y_i} \frac{\partial y_i}{\partial w}. \tag{1}$$

Finding the gradient $\partial I/\partial w$ can be split into the sample wise computation of the information forces $\partial I/\partial y_i$ and the adaption of the parameters $\partial y_i/\partial w$.

The second part is the simple one, since there exists a number of suitable transformations g, e.g. linear transformations or neural networks like Radial Basis Function Networks [1] or Multi Layer Perceptrons [7]. The only requirement is that they have to use the gradient information $\partial y_i/\partial w$ to adapt their parameters. All following examinations are limited to the linear transformation case, because this allows easy comparison with PCA or LDA and a visual inspection

of the results is possible as well. The parameters w that have to be estimated are all elements of the linear projection matrix W. The equation for the linear transformation is given by

$$y_i = W^T x_i. \tag{2}$$

The size of W is d_x times d_y with $d_x > d_y$ where d_x is the number input of dimension in X and d_y is the dimensionality of the subspace. Furthermore, W is assumed to be orthonormalized.

The calculation of the information forces $\partial I / \partial y_i$ is computationally more demanding. The straightforward approach would be to use the well known *Mutual Information*

$$I(Y,T) = \int_y \int_t P(y,t) \log \frac{P(y,t)}{P(y)P(t)} dtdy \tag{3}$$

to evaluate the correspondence between the transformed data and the target values. But due to the associated problems of estimating this criterion in high dimensional spaces, Torkkola proposes a non-parametric estimation based on *Quadratic Mutual Information* I^2

$$I^2(Y,T) = \int_y \int_t (p(y,t) - p(y)p(t))^2 dtdy \tag{4}$$

and kernel density estimation with Parzen windows. Application of the binomial formula splits equation 4 in three parts which are interpreted as information potentials and the derivatives as information forces.

$$\frac{\partial I^2}{\partial y_i} = \frac{\partial V_{IN}}{\partial y_i} + \frac{\partial V_{ALL}}{\partial y_i} - 2\frac{\partial V_{BTW}}{\partial y_i} \tag{5}$$

V_{IN} represents the "attractive potential" of all samples with the same/similar target value, V_{ALL} is the same but for all samples, and V_{BTW} is the "repulsive potential" (negative sign) between samples of different target values. The derivatives show the direction each sample has to move to maximize the objective function. The actual computation of these terms is reduced to interactions between all pairwise samples using Gaussian kernel density estimates. The reader is referred to [1] for the details that are omitted here.

In Algorithm 1 the procedure for one adaption step is given. These steps are repeated until convergence of the parameters w.

3 Image Data Processing

According to Torkkola [1], the previously described system is suitable for small input dimensions, but higher dimensions can be problematic. On one hand, image data is intrinsically high dimensional, because each pixel position is considered an input. On the other hand, treating each pixel as an independent input channel neglects the fact that neighbor pixels from the camera are dependent on each

Algorithm 1 Maximal Mutual Information Adaption Step

Input: current transformation W_t, the input data X and the target values T
Output: new transformation W_{t+1}

$Y = g(W, X) = W^T X$ // computation of the transformed data
$\frac{\partial I^2}{\partial y_i} = \frac{\partial V_{IN}}{\partial y_i} + \frac{\partial V_{ALL}}{\partial y_i} - 2\frac{\partial V_{BTW}}{\partial y_i}$ //estimation of the different forces
$\frac{\partial Y}{\partial W} = X^T$ //The gradient of the linear transformation matrix W
$W_{t+1}^{'} = W_t + \alpha \frac{\partial I}{\partial W} = w_t + \alpha \sum_{i=1}^{N} \frac{\partial I}{\partial y_i} \frac{\partial y_i}{\partial W}$ //Adaptation step
$W_{t+1}^{'} = \textsc{GaussianFilter}(W_{t+1}^{'})$ //Supplemental step for images, see Sec. 3
$W_{t+1} = \textsc{GramSchmidt}(W_{t+1}^{'})$ //Orthonormalization step to ensure $W^T W = \mathrm{I}$

other. We assume that informative parts of the image are not defined at pixel level, but by a more general, arbitrary shaped region, that is approximated on the pixel level. Thus, it is very unlikely that neighboring pixel have a rank different information content.

To cope with this problem and forcing the filter to consider these neighborhood dependencies, we introduced an additional step into the algorithm. After computing the new filter according to the gradient information and before the orthonormalization step, we perform a smoothing with a Gaussian filter in the two dimensional image space on the filter mask. This does not only distribute information between neighbor input dimensions, but increases stability and convergence speed, because the algorithm finds smooth solutions. An additional benefit is the obvious reduction of measurement noise in the observations.

This approach is not suitable for images only, but every continuous domain that is sampled and approximated at certain points and has a clear neighborhood definition.

We used 1.440 small images of the size 40x32 pixels which equals 1.280 input dimensions for each sample. All images are flame pictures taken from a coal burning power plant. An example image with a higher resolution is shown in Fig. 2. The respective targets are measurements like the nitrogen oxides (NOx), carbon monoxides (CO) or excess air (O2) produced by the combustion. We used Multi Layer Perceptrons as function approximators and evaluated the performance of different instances of the ITL framework with different parameters like the dimensionality of transformed data, and compared them with results from PCA based transformations.

The first tests were made with a visual examination of the resulting filter masks, similar to the well known eigenfaces. Images with obvious structures are considered "stable" solutions, while "unstable" results are characterized by high frequent noise and no structures in the filter masks. See Fig. 3 for examples on a higher resolution (134x100 pixels). Interestingly, if stable solutions were found, they tend to be similar to each other, besides differences in the sign of the filter masks, which relates to the same axis but the opposite direction. More discussions on this topic are following in Sec. 4.

Different initializations for the transformation parameters w result in the clear preference for PCA or LDA, since randomly initialized filters tend to pro-

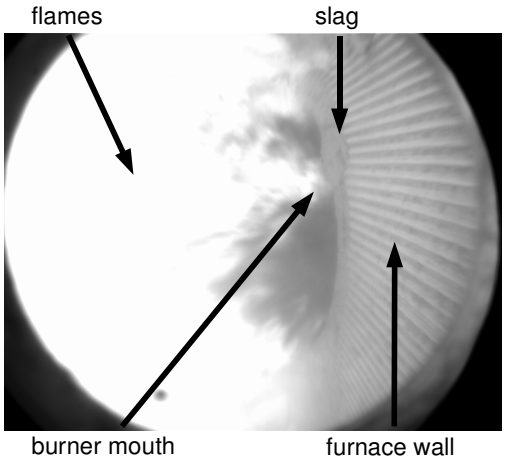

Fig. 2. An example of a black and white image taken from the furnace of a coal fired power plant. Clearly visible is the furnace located wall on the right. Roughly in the middle of the images is the burner mouth were coal dust is inserted to the furnace and ignites. Around this area on the wall, slag (molten ash) is visible.

duce unstable results. Hence the MMI method is more of an objective driven refinement for these plausible starting guesses.

The possible dimensionality of the reduced feature space d_y is greatly dependent on the number of available samples. This makes sense with respect to the curse of high dimensionality, because the higher the dimensionality, the more difficult it is to estimate the required probability distributions. For the presented setup of data we noticed two things: First, the bigger d_y, the more it deviates from a PCA initialization. Second we observed that the breaking point, where it switches from stable subspace transformations to unstable results, is between $d_y = 4$ and $d_y = 5$. By doubling the number of samples to 3.600, we get stable results in the five dimensional subspace, but $d_y = 6$ and higher remain unstable.

Further experiments where conducted with images subscaled to an even smaller size of 10 x 8 pixels per image. The reductions of the input dimensionality d_x does not improve the results considerably. On the other hand, using images with 160x120 pixels decreased stable results to 3 dimensions. This is due to the linear connection between d_x and the number of parameters w compared to the exponential influence of d_y as discussed above.

The next experiments are conducted to test whether the MMI subspace transformation yields any improvements compared to PCA-based *eigenflames*. Taking the previous results into account, the target dimensionality is limited to $d_y <= 3$ and the MMI subspace search started with an PCA initialization.

The results clearly demonstrate the benefits of the MMI approach. The approximation errors are smaller or at least in the same magnitude of the PCA-based approach. By adding more channels, the PCA can achieve similar results to the MMI transformation, but there is always the need of additional dimen-

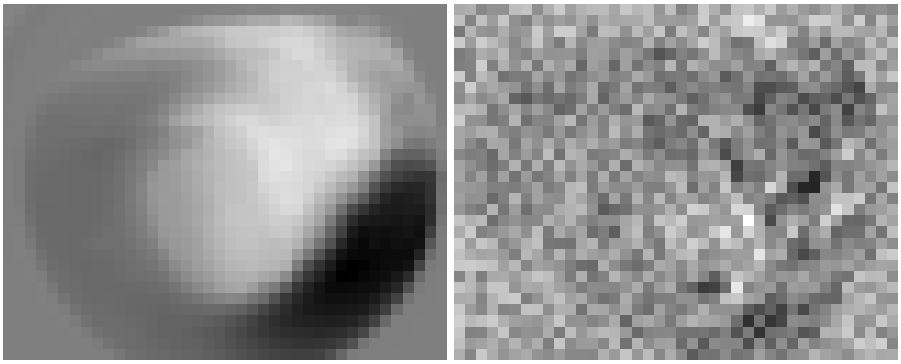

Fig. 3. (Left) A stable filter mask. **(Right)** An unstable one. Both subspace transformations are the results of the optimization procedure described in this section and depict the first dimension of the new subspace. The white areas are coding positive values, the black regions negative ones, while the gray areas are near zero and thus unimportant, like the round margin in the left image. One important fact to remember is, that the sign does not tell anything about the importance of this pixel, while the absolute value does. This kind of visualization is comparable to Eigenflames produced by PCA, besides in this case it doesn't depict the variances in the data, but the information.

Table 1. Comparison of the same MLP trained with PCA subspace features or MMI subspace features respectively for three different targets. All prediction errors are the MSE from an independent test set. The high level of noise present in the data leads sometimes to the effect of increasing errors when providing additional input features.

	Prediction Error for CO		Prediction Error for O2		Prediction Error for NOx	
d_y	PCA	MMI	PCA	MMI	PCA	MMI
1	3.11	3.07	0.90	0.24	28.88	25.99
2	3.33	2.43	0.25	0.29	35.50	25.00
3	4.07	2.66	0.22	0.28	27.65	30.26

sions to represent the information. Hence, we conjecture that the MMI method is able to compress the relevant information better than the PCA *eigenflames*.

One negative aspect concerning the MMI approach are the computational costs associated with the density estimation and gradient computation. While PCA is fast to compute, MMI takes a lot of time (which is mainly dependent on the number of input images used). For several thousand images it can easily take one or two hours to obtain the filter masks. Hence, the MMI methods can be applied only if there are no hard time constraints.

To conclude this section, the experiments show that it is beneficial to use the MMI system to improve PCA based subspace transformation for image data.

4 Adaptive Feature Transformation

It is assumed that the presented system is used as a preprocessor for a controller or function approximator which is able to handle slow adaptations itself. The goal for the adaptive feature extraction system is to provide similar features if the underlying process is in a similar state, and different features in different process states.

There are several configurations of the subspace transformation parameters w possible that achieve a maximal value with respect to the optimization goal of maximizing the Quadratic Mutual Information even for the optimistic case of a single, global maximum. For example, a scaling of the matrix W with a non-zero scalar will not change the information content of the results. All but two of the possible solutions are eliminated during the orthonormalization step. This step restrains all configurations in the parameter space to the hyper unit sphere. The two remaining valid solutions are w^* and $-w^*$. These transformations obviously contain the same information, since the only discriminating feature between data transformed by the two filter is the inverted sign. This behavior is not desired since the same state can yield two different subspace transformations that produce opposite transformed data, which are completely different to the system using the transformed features.

If the process is stationary, this problem can be overcome by the use of a suitable similarity measure to compare the old filter configuration w_{old} to w_{new} and $-w_{new}$ accepting the better match. But it is quite hard to define good similarity measures and thresholds if the process is transient. The most obvious work around is a different starting initialization. Instead of starting from the PCA subspace, it is possible to use the previous MMI subspace as initialization point. Assuming that the transient process changes are slow, compared to the adaptive updates of the filter, these changes will yield slow changes of the relevant feature areas. Thus, the subspace transformations will be similar to each other, which justifies the use of the previous solution as a starting point.

The actual adaptivity can be achieved on different time scales. One possibility is to adapt the current filter into the new one after a few measurements, using the techniques described in [1](Appendix A) where not the whole available data is used for a adaption step but only a small subset. The extreme case is the use of two samples. Torkkola draws them randomly, while for an online system these samples are the last measurements. For those samples one adaption step is applied (see Algorithm 1).

For applications with very noisy measurements, this may introduce the problem that the systems tries to adapt to the noise, rather than the underlying process changes. Hence, slower timescales change the procedure to collecting a certain amount of data before performing a batch update of the filter.

For the online application of the system of the power plant, we are interested mainly in very slow changes induced by wear and tear of the furnace or coal type changes. There are other changes on a much faster timescale, but they are even harder to detect, due to the presence of a high measurement noise. For the experiments presented here, a daily batch update was used. We used the

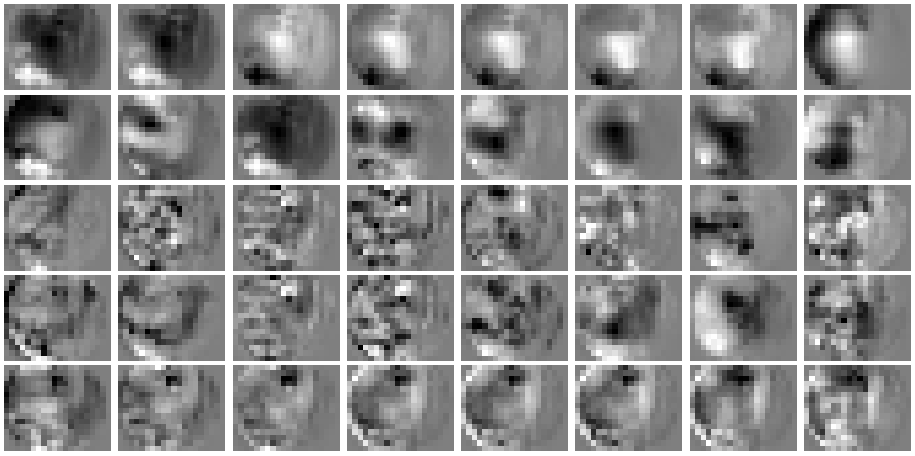

Fig. 4. Each of the columns represents the results of the MMI adaption process for the flame images and always shows the first component of the new subspace. The target values used for the training process of the LDA and MMI are the corresponding nitrogen oxides (NOx) measurements. The differences in each row are the used subspace transformations. **First row:** PCA. **Second row:** LDA. **Third row:** MMI initialized with the current PCA result. **Fourth row:** MMI initialized with a global PCA result. **Fifth row:** MMI initialized with the previous MMI result. The most interesting observations are the changes over time (from column to column), since smaller changes are desired.

collected data of five elapsed days for training purposes, and the most recent day as test set. We used this data to form PCA, LDA and MMI subspaces for eight consecutive days. For the MMI method we employed three different initialization points. First, we used the result of the PCA on that time frame for this purpose. Second, a fixed *eigenflames* subspace transformation calculated over the complete data was used, and third, the previous MMI result was used for the initialization.

Some results of these experiments are shown in Fig. 4. The PCA results (first row) are the most stable ones over time, the variance in the data over time is similar. But here again is the possible pitfall of the sign inversion problem between column 2 and 3. The LDA results (second row) identify big connected regions, but the shapes are completely different each day. Independent of the initialization, all MMI results share the tendency to produce less homogenous regions. The MMI results based on the the PCA initializations (third and fourth row) behave similar to the LDA subspace transformations, they are different each day. Using the previous MMI subspaces as starting points (last row) yield very useful but adapting filter masks.

These experiments show that the initialization with the previous MMI results is the most promising way to handle the adaptation task in a changing environment.

5 Conclusion and Future Work

Our experiments using the MMI feature subspace transformations for image data processing show that the approach is indeed useful, but has its limitations. The information extracted is either more informative for a classifier than a PCA-based subspace, or at least it is possible to compress the same information into a lower dimensional subspace than PCA. But to achieve stable results the use of PCA as a initialization is required anyway, so the MMI is in practice a objective driven refinement of the results obtained by PCA or LDA.

These positive results only hold true for rather low dimensional subspace constructs. If the desired transformation projects into a still high dimensional space, the MMI approach will get stuck at a local minimum very soon or venture into directions were stable solutions are hard to find by gradient descent. In these cases the use of LDA or PCA is superior to the MMI method.

The stepwise gradient estimation of the MMI subspace is an advantage for an adaptive online system. It allows the use of previous solutions to estimate a similar subspace which captures at least some changes of the underlying process without a complete redefinition of the channels in the new subspace.

Possible directions for future work include the investigation of the extension to nonlinear transformations like neural networks in the image domain. The adaptive changes of the subspace transformations focus on finding that subspace which is most important to the tasks at hand are engineered from the practitioners point of view. Hence investigating the connection of our proposed system to biological inspired, attention-based systems would be an interesting venue, too.

References

1. Torkkola, K.: Feature Extraction by Non Parametric Mutual Information Maximization. *Journal of Machine Learning Research* **3** (2003) 1415–1438
2. Duda, R., Hart, P., Stork, D.: *Pattern Classification.* New York, Wiley (2001) 2nd ed
3. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis.* New York, Wiley (2001)
4. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems* **13**. MIT Press (2001) 556–562
5. Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**. (1936) 179–188
6. Principe, J., Xu, D., and Fisher, J.: Information theoretic learning. In Haykin,S., editor, *Unsupervised Adaptive Filtering.* Wiley (2000) 265–319
7. Torkkola, K.: Nonlinear feature transforms using maximum mutual information. *Proc. Int. Conf. Neural Networks* (IJCNN). (2001) 2756–2761